Paper submitted at the European Population Conference
Padua, 24-27 June 2020

The curse of the plateau. A skeptical view on mortality estimation at extreme ages

Carlo G. Camarda
Institut National d'Études Démographiques
carlo-giovanni.camarda@ined.fr

Abstract

Importance of describing mortality at the limits of life span has lately lead to relevant as well as controversial articles. Whereas considerable effort has been made in collecting data and estimating models on oldest-old individuals, testing statistical confidence about eventual conclusions has been largely neglected. How certain can we be saying that risk of dying increases, levels out, or paradoxically decreases over age 105? Can we detail particular mortality age-patterns at those high ages? In this paper, it is shown that very little can be affirmed when we venture in describing mortality at extreme ages. Instead of analyzing actual data, we perform a series of simulation studies mimicking actual scenarios. By knowing the true underlying age-patterns, we generate lifetimes which are either fully observed or censored/truncated from controlled mechanisms. Our findings are thus robust with respect to factors such particular observation schemes, heterogeneity and data quality issues. Given sample sizes currently available and levels of mortality experienced in present populations, we show that before age 110 only a gompertzian increase of mortality can be eventually detected. Afterwards a plateau will be regularly recognized as the most suitable pattern, regardless the complexity of the true underlying mortality.

Keywords: Extreme ages; Gompertz model, Likelihood-ratio test; Mortality plateau; Simulations; Smooth hazard

Null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation.

Ronald Fisher (1935)
The Design of Experiments
Oliver and Boyd, p. 18

1 Introduction

Knowledge about the shape of mortality at very high ages, here above 105, is essential to test possible evolutionary theories and to understand whether humans experience a limit in the increasing risk of dying. However, estimating mortality at extreme ages poses two main related issues: we deal with small sample sizes that experienced high mortality. Consequently, we work with populations that might render any study fallacious. Despite these intrinsic limitations, several researches attempted to shed light on the age-pattern of mortality at very high ages, with mixed outcomes.

A crucial aspect of this type of data is the reliability of the reported age: we commonly deal with old vital registries as well as with age misreporting and exaggerations. An example of collective effort to provide thoroughly validated information on individuals who attain extreme ages is the International Database on Longevity (2019, IDL). Since twenty years, this research consortium have being collecting (and analyzing) data on the so-called *supercentenarians* (who has lived ≥ 110 years) and then on the larger group of *semi-supercentenarians* (≥ 105).

Concerning the analysis of these and other analogous mortality datasets, and at the risk of simplifying the scene, we could identify two antithetical conclusions. On the one hand, several studies, based on individual data, have pointed out the existence of a plateau when individuals reach extreme ages. On the other, signals of a continuous increasing of mortality have been acknowledged by other researches. In the former group, we can recognize the studies by Gampe (2010) and Rootzén and Zholud (2017) that, using IDL data, show a constant risk of dying after age 110. Barbi et al. (2018b) present an analysis of Italian data and indicate an essentially constant mortality beyond age 105. Explicit in its conclusions, doubts have been raised on this last study and a stimulating debate have prompted (Barbi et al., 2018a; Beltrán-Sánchez et al., 2018; Camarda et al., 2018; Gavrilov and Gavrilova, 2019; Medford, 2018; Milholland et al., 2018; Newman, 2018a,b; Olshansky and Carnes, 2018; Wachter, 2018).

Conversely, absence of a mortality plateau have been found by Gavrilova et al. (2017) that, using IDL data, have that mortality after age 110 years do not stay constant. An analysis on French *semi-supercentenarians* revealed that a constant mortality level beyond age 105 can be statistically rejected (Dang et al., 2019). Using individual Dutch data, Einmahl et al. (2019) claimed a statistical evidence about an upper limit to the life and therefore to the absence of a mortality plateau.

With the objective of better understand human mortality at extreme ages, age trajectories and their development over time, another strand of research have employed aggregate data. For instance, a steady increase in the risk of dying at very high ages have been advocated (Gavrilov et al., 2017; Gavrilova and Gavrilov, 2011, 2015), and analysis of time trends in maximum attained ages have been used to state that human lifespan has reached its limit (Dong et al., 2016; Milholland et al., 2017; Vijg and Le Bourg, 2017). However, doubts on

data quality and statistical approaches in these studies have been also raised (Brown et al., 2017; de Beer et al., 2017; Hughes and Hekimi, 2017; Keiding, 2018; Lenart and Vaupel, 2017; Rozing et al., 2017) and other studies have reached different conclusions (Antero-Jacquemin et al., 2015; Lenart et al., 2018; Rau et al., 2017).

In this study, we make a step back and re-formulate the main questions. Before disputing whether the risk of dying is increasing or stalling and consequently speculating in terms of biological theories and cure for aging, it is necessary to figure out how confident we are about any eventual outcome. We are thus not interested in the actual shape of mortality at high ages, but rather in the amount of reliance expected when claims are made about longevity above age 105.

In order to pursue with this aim, we free our analysis from any possible issue actual datasets bring along and work on simulated data only. Specifically, we generate fully observed lifetimes which are eventually censored and truncated using known mechanisms. Consequently specific observation schemes will solely influence the estimated uncertainty in the results. Moreover sample selections is not considered and they thus cannot beset our outcomes. Only random fluctuations due to sample size will affect estimated values and heterogeneity in frailty cannot be an explanation for specific conclusions. Most importantly, we have a complete knowledge of the underlying risk of dying and therefore we are able to exactly test whether a specific model is able to properly describe the true latent mortality.

In the following two simulation studies are considered. First we assume lifetimes as a realization of a Gompertz law, that is a exponential increase of mortality over age. We use different starting levels of mortality, values of rate-of-aging and sample sizes. By varying these factors, we test the capability to discriminate between true underlying pattern and simpler exponential distribution, which assumes a constant level or mortality, i.e. a plateau. By incorporating censored and truncated mechanisms, we show how these factors are simply decreasing the confidence about any finding.

Still we are in a dichotomous world in which either we accept a Gompertz or a plateau at extreme ages. In order to overcome this condition, a second simulation setting will free the estimated mortality age-pattern from any rigid structure. Starting from true Gompertz distributions with different sample sizes and observation schemes, a smooth risk of dying will be estimated to assess if patterns might be identified for specific age ranges.

To better mimic actual scenarios, we employ true values which are about those obtained by estimating similar models on actual data (Barbi et al., 2018b; Dang et al., 2019; Gampe, 2010; Gavrilova and Gavrilov, 2011; Rau et al., 2017). However we consider a much larger spectra of possible age-patterns and test their plausibility estimating highest attained age as in Thatcher (1999).

The remainder of this paper is structured as follows. Section 2 presents basic assumptions and describes survival analysis concepts which lay the groundwork for further steps. We illustrate how to simulate lifetimes from parametric models as well as censoring and truncation mechanisms. In Section 2.2 a penalized version of the likelihood for Piecewise constant hazard model is used to estimate smooth hazards. Outcomes for both parametric and non-parametric approaches are given in Section 3. In the first part we present how often and in which scenarios we are able to retrieve from simulated data the true underlying model using classic likelihood-ratio test within a parametric setting. In Section 3.2 with no assumption on the mortality age-pattern, we show how challenging is to capture the true Gompertz model at extreme ages. Finally, we draw our conclusions in Section 4.

2 Simulating and estimating survival models

When the interest is the description of the risk of dying in a certain population, survival analysis provides statistical tools for analyzing individual data. As in a regression setting, we have a dependent variable or response which is, in our context, the waiting time until death. Let denote by T the non-negative continuous random variable representing time-to-death. We can describe T with a probability density function $f(t)$ and a cumulative distribution function $F(t) = Pr\{T \leq t\}$. However, it is often more convenient to work with associated survival and hazard functions:

$$\begin{aligned} S(t) &= Pr\{T > t\} = 1 - F(t) = \int_t^\infty f(x)dx \\ h(t) &= \lim_{dt \to 0} \frac{Pr(t < T \leq t + dt | T > t)}{dt} \,. \end{aligned} \tag{1}$$

Whereas $S(t)$ gives the probability of being alive at time t and it is the complement of $F(t)$, the hazard function can be interpreted as the instantaneous death rates and demographers commonly labeled it "force of mortality". Relationships between these functions can be easily derived and sum up by the following equations:

$$h(t) = \frac{f(t)}{S(t)} \quad , \quad h(t) = -\frac{d}{dt} \ln S(t) \quad , \quad S(t) = \exp\left[- \int_0^t h(u)du \right] \tag{2}$$

For further details, see Klein and Moeschberger (2003).

2.1 Parametric models

The simplest possible survival distribution is obtained by assuming a constant risk over age:

$$h(t; a) = a \geq 0 \,. \tag{3}$$

Commonly labeled as *Exponential*, this distribution plays a central role in survival analysis. In our context, finding a plateau means that mortality at extreme ages is well described by (3).

Following Gompertz (1825), we can define a slightly more complex distribution by assuming an exponential increase of mortality over age:

$$h(t; a, b) = ae^{bt} \,, \tag{4}$$

where $a, b \geq 0$ are parameters used to describe the starting level of mortality and rate-of-aging, respectively. Often applied to describe distribution of adult lifespans, the Gompertz law can be considered a generalization of the Exponential distribution with $b = 0$. In other words, if a dataset can be better described including the parameter b, a mortality plateau should be dismissed as hypothesis.

In our simulation settings, individual times-to-death will be generated by a Gompertz distribution with given parameters a and b using the inverse transform sampling method (Devroye, 1986). Given (4) and inverting the associated cumulative distribution function, N sample lifetimes are computed as follows:

$$t_i = \frac{1}{b} \ln \left[1 - \frac{b}{a} \ln (1 - u_i) \right] \,, \quad i = 1, \ldots N \tag{5}$$

where $u_i \sim U(0,1)$.

A common feature in actual supercentenarian datasets are censored and truncated observations. In our simulation setting, we only account for the most common right censoring and left truncation, both considered non-informative. In the former scheme we only know that a supercentenarian is still alive at a given time. In the latter situation, only supercentenarians who survive a sufficient time are included in our simulated dataset. Other observation schemes (left censoring, right truncation and interval censoring) may be included making the whole estimation procedure statistically more complex and without modifying our final conclusions. Indeed, their inclusion will lead to higher uncertainty in the results. Our outcomes can thus be considered more conservative with respect to what one would obtain by analysing actual datasets with all possible censoring and truncation schemes.

To simulate survival data with censoring, we need to model the hazard function for time to censoring. In the following, we opt for a simple Exponential distribution: the risk of being censored is constant over age and independent from the risk of dying. For ease of presentation we label by ν the parameter for the censoring times c_i. As in (5), we proceed for the Exponential distribution and obtain:

$$c_i = \frac{-\ln(1-u_i)}{\nu}, \quad i = 1, \dots N \tag{6}$$

where u_i are random number from a Uniform distribution. The observed exit time is the minimum of the censoring and lifetimes: $y_i = min\{t_i, c_i\}$. Moreover an indicator variable, called δ_i, will be generated stating whether observation terminated by death ($\delta_i = 1$) or by censoring ($\delta_i = 0$).

Left truncation is simulated by assigning to randomly selected $L \leq N$ individuals an entry time w_i which is generated by a Uniform distribution with minimum and maximum such that $0 \leq w_i \leq y_i$. For all other $N - L$ observations, w_i will be equal to zero, i.e. age 105.

Given a set of simulated triplets $(w_i, y_i, \delta_i), i = 1, \dots, N$, we employ the maximum likelihood estimation to obtain the parameters from a specific model and, for convenience, we work with the natural logarithm of the likelihood function. We maximize the following function:

$$\ell(\boldsymbol{\theta}) = \ln \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{N} \{\delta_i \ln h(y_i; \boldsymbol{\theta}) + \ln S(y_i; \boldsymbol{\theta}) - \ln S(w_i; \boldsymbol{\theta})\} \tag{7}$$

where $\boldsymbol{\theta}$ identifies either the constant hazard a, or the Gompertz set of parameters $\boldsymbol{\theta} = (a, b)$. Maximization is obtained numerically with `R` routines devised by Jackson (2016).

Since the Exponential can be nested within the Gompertz distribution, we use the likelihood-ratio test for comparing the goodness of fit of these two demographic models. Let denote by $\mathcal{L}_E(\hat{a})$ and $\mathcal{L}_G(\hat{a}, \hat{b})$ the maximized likelihoods under the Exponential and Gompertz model, respectively. The ratio of these two quantities

$$LR = \frac{\mathcal{L}_E(\hat{a})}{\mathcal{L}_G(\hat{a}, \hat{b})} \tag{8}$$

is bounded between 0 and 1, and it expresses how many times more likely the data are under one model than the other. Under certain regularity conditions, minus twice the log of the likelihood ratio has approximately a χ^2 distribution with degrees of freedom equal to the difference in the number of parameters between the two models, 1 in our context:

$$-2 \ln LR = 2 \ln \mathcal{L}_G(\hat{a}, \hat{b}) - 2 \ln \mathcal{L}_E(\hat{a}) \to \chi^2_1 . \tag{9}$$

The knowledge about the asymptotic behavior of this test can be used to compute a p-value and to provide a decision rule in selecting the model. For instance, if we reject the plateau hypothesis when the p-value is smaller than 0.05, it means that we consider tolerable a maximum of 5% probability to commit the error "reject a mortality plateau hypothesis when is actually true".

2.2 Smooth hazard

Moving away from a parametric setting, Piecewise constant hazard models are a more flexible option for describing hazard function. In general, we partition the age axis into m intervals and we assume the hazard to be constant within each interval. Widely used in demography when prior knowledge of the underlying hazard function is weak, this model can be estimated using Poisson regression (Holford, 1980). Two limitations are evident in this approach: estimated hazard function will be discontinuous and there is a degree of subjectivity in the choice of the breakpoints.

In order to overcome both issues simultaneously, we propose to intentionally over-parametrize the basic model with a large m and to restrict, via a penalty, all redundant features for achieving a *wisely* parsimonious description of the hazard. Specifically, we divide the age axis into narrow bins of equal length δ and we compute two m-dimensional vectors $\boldsymbol{y}$ and $\boldsymbol{e}$ containing, in the m bins, the numbers of deaths and the total exposure times experienced by all N individuals. We can estimate the constant hazard h_j in each bin j by assuming that y_j are Poisson distributed counts with expected value $\mu_j = h_j e_j$.

Without additional assumptions, h_j simply corresponds to death rates for each bin. Consequently estimated hazard trajectories will be remarkably jagged when m gets larger (i.e. δ smaller), also taking into account that we deal with relatively few deaths. We thus decide to enforce smoothness over age to obtain a reliable description of the force of mortality without assuming a strong structure of its pattern. This approach also allows us to find out trajectories for specific age ranges without imposing a general law of mortality for all ages.

Within a Poisson regression, logarithm is used as link-function and estimation is obtained for the log-mortality, $\boldsymbol{\eta} = \ln \boldsymbol{\mu}$. Smoothness can be achieved by adding a difference penalty in a standard Poisson regression setting (Eilers and Marx, 1996). Let denote by $\boldsymbol{D}_d$ the difference matrix of order d, we can adopt a penalized version of the iteratively reweighted least squares algorithm to estimate a smooth hazard:

$$(\tilde{\boldsymbol{E}} + \lambda \boldsymbol{D}_d' \boldsymbol{D}_d)\tilde{\boldsymbol{\eta}} = \boldsymbol{y} - \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{E}}\,\tilde{\boldsymbol{\mu}} \tag{10}$$

where $\tilde{\boldsymbol{E}} = \texttt{diag}(\tilde{\boldsymbol{\mu}})$ and the tilde-symbol denotes current approximations to the solution. The smoothing parameter λ regulates the trade-off between goodness-of-fit and effective dimension used into the model. On the one hand higher values will lead to higher penalty term and, consequently, smoother hazard. On the other, $\lambda = 0$ results to a straightforward Poisson estimation, i.e. computation of m death rates for each bin. It is noteworthy that the number of intervals (i.e. the size of δ) is immaterial and it does not modify the final $\hat{\boldsymbol{\mu}}$, as long as m is selected sufficiently large to capture all important features in the data. Here we use $\delta = 0.05$ which corresponds to 20 intervals each year of age, a h_j every 18.25 days.

The shape of the hazard when $\lambda \to \infty$ will be a polynomial sequence of degree d at the log-mortality ($\boldsymbol{\eta}$) level. Here, we decide to present results for both $d = 1$ and $d = 2$ and hence, for really large lambda, we would obtain either a constant or a gompertzian force of mortality over age, respectively. In other words, the choice of d could be interpreted as a

prior idea of the underlying force of mortality when no additional information and data are provided.

The advantage of this penalized approach lies also in the fact that we can tune the whole model by changing λ only. In the following the smoothing parameter will be selected by minimizing the Akaike Information Criterion (AIC) which, at the cost of wiggling estimated hazard, will not oversmooth important features in the data. For a given λ the AIC is computed as follows:

$$AIC = DEV + 2\,ED \tag{11}$$

where the deviance $DEV = 2\sum_j y_j \ln(y_j/\hat{\mu}_j)$ measures the goodness-of-fit in the model and the effective dimension is $ED = \mathtt{tr}[(\hat{\boldsymbol{E}} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\hat{\boldsymbol{E}}]$. The minimum of the AIC is found by a grid-search over a wide range of possible λ.

Since we are in a regression Poisson setting, diagnostics and measures of uncertainty are easily obtained in the proposed penalized hazard estimation. Specifically, the covariance matrix of the estimated log-mortality, $\boldsymbol{\eta}$, (for fixed λ) is given by

$$\boldsymbol{C}_{\boldsymbol{\eta}} = (\hat{\boldsymbol{E}} + \lambda \boldsymbol{D}'\boldsymbol{D})^{-1}\,. \tag{12}$$

The covariance of the fitted hazard $\boldsymbol{\mu} = e^{\boldsymbol{\eta}}$ follows as $\mathtt{diag}(e^{\hat{\boldsymbol{\eta}}})\,\boldsymbol{C}_{\boldsymbol{\eta}}\,\mathtt{diag}(e^{\hat{\boldsymbol{\eta}}})$. The square root of the diagonal of this last matrix provides the standard errors for the estimated hazard which are then used to construct the associated confidence intervals.

Alternative and more complex approaches for obtaining a smooth estimation of the force of mortality have been tested. For instance, we combined penalties of first and second order as in Eilers and Goeman (2004) to merge Exponential and Gompertz models in the penalty term, and we adopted an adaptive smoother as in Wood (2006) and Krivobokova et al. (2008) allowing for varying smoothness over age. Outcomes from these methods are substantially equal to the proposed penalized approach in terms of fitted hazard and associated confidence intervals.

3 Plateau, or not plateau, is that the question?

In the following we will present two simulation studies. First we simulate lifetimes from a Gompertz distribution and we test if and when, by changing sample size and true parameters, we are able to discriminate between the true underlying model and the simpler constant hazard model. Then we will show nonparametric estimations of the hazard for various datasets, again generated from a Gompertz model.

3.1 Discriminating Constant and Gompertz hazard

For simulating lifetimes from a Gompertz distribution as in (5), we need to select both sample size (N), starting level of mortality (a) and rate-of-aging (b). Since our focus is on the significance of the b parameter, we take only four possible $a = (0.5, 0.6, 0.7, 0.8)$. However, the range of these values contains all possible starting levels of mortality currently estimated at age 105 (Barbi et al., 2018b; Dang et al., 2019; Gampe, 2010; Gavrilova and Gavrilov, 2011; Rau et al., 2017) and more than 70% of the rates at 105 provided at aggregate level in the Human Mortality Database (2019) for all countries in the last 50 available years.

Rate-of-aging in a Gompertz distribution measures the relative derivative with respect to age of the force of mortality. Effects of a small change in b is very large in the final mortality age-pattern, especially in small population (Witten and Satzer, 1992). However, the marginal

effect of b on the probability of being selected is linear in terms of percentage changes. Therefore we used a detailed series of true b and in a log-scale: $\log b = (-2, -1.95, -1.9, \ldots, -1)$. These values correspond to an unequal-spaced series from 0.01 to 0.1. Figure 1 shows all underlying mortality age-patterns assumed in our simulation study. We cover a wide range of possible trajectories, i.e. all combinations of high/low mortality levels with slow/fast mortality changes over age.

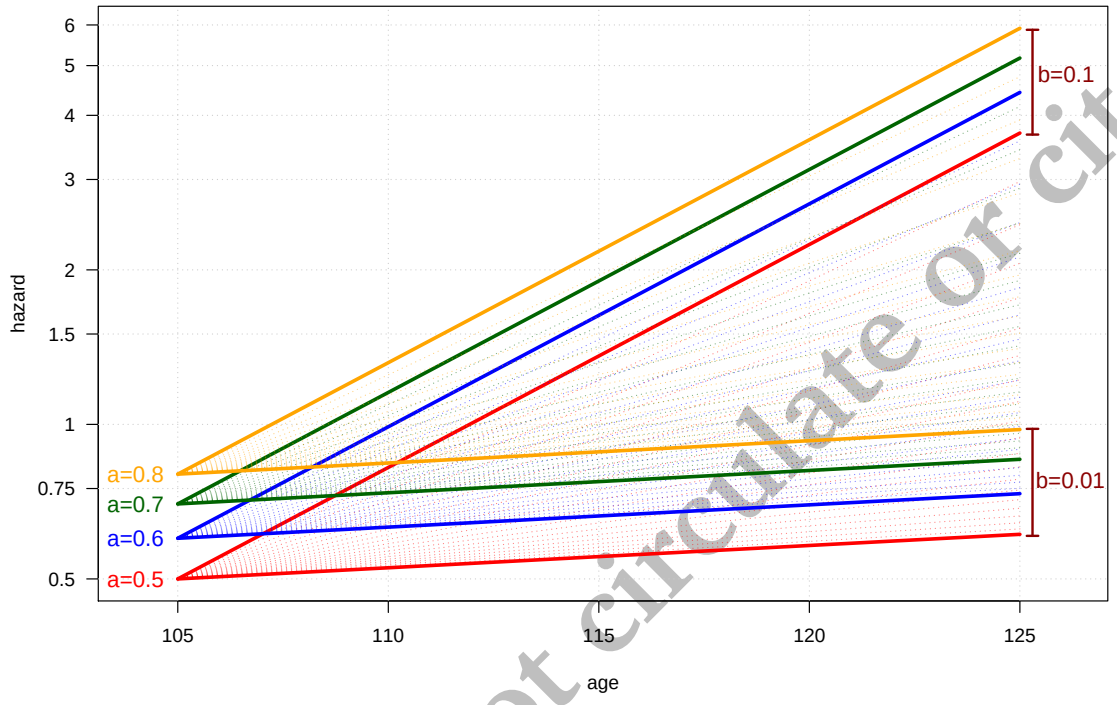

Figure 1: A schematic overview of all possible underlying mortality age-patterns used in the simulation study.

Sample size is another critical aspect in our context. Hence we simulate using a detailed series of N. Specifically we will have datasets with $N = (1, 2, 3, \ldots, 20) \times 10^3$. We show that we can hardly conclude from our smallest N, and the largest value (20,000) is more than the total number of observations currently available in the International Database on Longevity (2019, IDL).

For all combinations of 20 possible N, 4 different starting levels of mortality and 21 b, we simulate 500 datasets. This implies that we simulate 840,000 independent datasets. For each dataset, we estimate both Exponential and Gompertz distribution and we test via a likelihood ratio if we could have rejected the null hypothesis (a mortality plateau). We use 5% as significance level of the test. In other words, for a given set of (N, a, b), we asses 500 times whether we could have retrieved the true underlying Gompertz distribution, or alternatively a simpler constant hazard would have been selected by the classic likelihood-ratio test.

To better acknowledge the effect of specific observation schemes, we compare outcomes from fully observed lifetimes with those obtained with right censored and left truncated observations. Specifically, we assume a constant hazard for time to censoring with rate equal to $\nu = 0.1$. This hazard produces between 7.3 and 20.6% of right censored observations

with relatively higher values when the underlying Gompertz hazard is lower. We also assume $L = N/10$ left truncated observations with entry times w_i uniformly generated between $y_i/10$ and $y_i/2$. Given these censoring and truncation mechanisms, we observe about 80-90% of the total exposure time we have with fully observed lifetimes.

For comparison, the most updated version of the IDL contains 22% of right censored and almost 99% of left truncated observations. Moreover, among the about 19,000 observations in the IDL above age 105, we have 78% of right truncated and less than 1% interval censored lifetimes. Another comparison with observed censoring and truncation mechanisms can be taken from Barbi et al. (2018b). They analyzed 3,836 individuals 105 and older with 33% right censored and 12% left truncated cases. As consequence, in our simulated datasets, we observe a larger number of details than we could detect in the most accurate mortality data for people at advanced ages.

Figure 2 presents the distributions of p-values for two instances of (N, a, b)-combinations, with and without fully observed lifetimes (top and bottom panels, respectively). If we have a dataset of 3,000 fully observed individuals with a level of mortality at 105 of 0.6 and an underlying rate-of-aging equal to 0.014, only 24.2% of the 500 simulated datasets are selected as Gompertz by the likelihood-ratio test with a significance level of 5% (top left panel). With larger sample size ($N = 5,000$) and a higher Gompertz parameters ($a = 0.7, b = 0.02$), this percentage increases up to 48.9% (top right panel). In both cases, we are below 50%: more then half of the times, we would not reject the hypothesis that b is equal to zero and consequently opt for a more parsimonious model, that is a mortality plateau.

Bottom panels in Figure 2 presents the same (N, a, b)-combinations, but right censoring and left truncation are incorporated in the simulation. Here the percentages are very low: in more than 95% of the 500 datasets a likelihood-ratio test will guide toward a constant mortality risk of dying. Cumulative distributions of the p-values depicted by the monotonically increasing lines in Figure 2 show how these percentages would change if the significance level varies. As expected, a smaller significance level (e.g. 1%) in discriminating between Gompertz and Exponential would produce lower percentages: a constant mortality pattern would be selected more often.

For all different model specifications we compute the percentages the true underlying Gompertz model is selected. Figure 3 and 4 present shaded-contour plots of these percentages when we consider fully observed lifetimes and censoring/truncation, respectively. In both figures, the four panels present 4 different starting level of mortality (a) adopted in the study. Sample size (N) and rate-of-aging (b) are on the horizontal and vertical axes, respectively. The color key drawn alongside the plot describes all possible percentages obtained in the study.

Regardless the values of the triplet (N, a, b), the true underlying model is a Gompertz. However, the relatively large areas of Figure 3 and 4 depicted by dark colors reveal that oftentimes we would select a simpler Exponential model, i.e. a mortality plateau. These percentages highly depends upon the values of (N, a, b). A higher starting level of mortality leads to lower probability to select the true Gompertz model. The higher the mortality at age 105, the fewer observations survive at older ages. Consequently, fewer observations will be available to properly estimate the underlying positive rate-of-aging.

Differently, larger values of b in the true hazard and a more numerous population are associated with higher chances to select the underlying Gompertz distribution, a held constant. The likelihood-ratio test will often reject the hypothesis of $b = 0$ if data were simulated with a relatively high rate-of-aging. The role of the sample size is obvious: a larger N helps to select the true underlying Gompertz model.

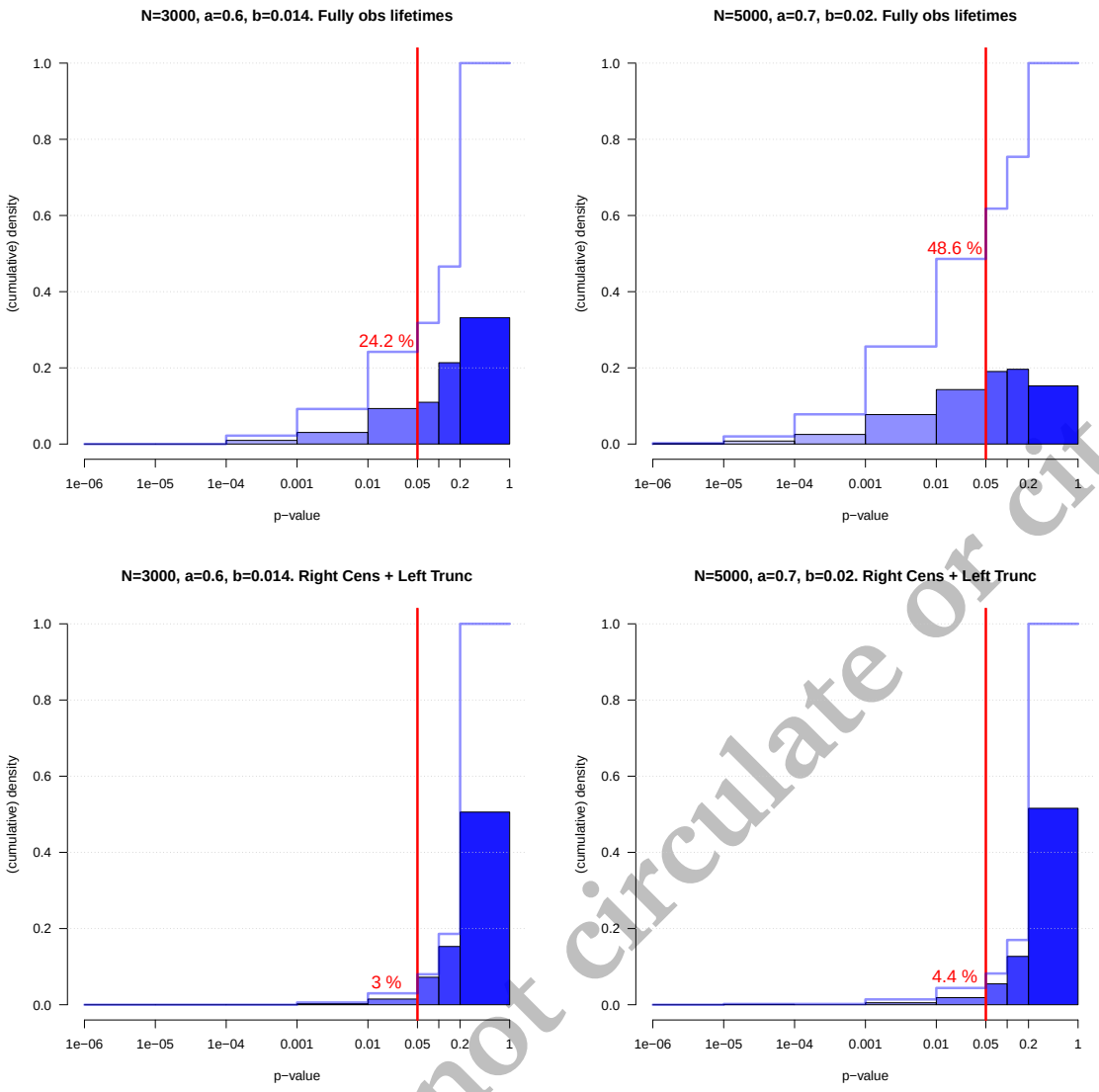

Figure 2: (Cumulative) Distribution of p-values from 500 datasets obtained by likelihood-ratio test on the hypothesis $b = 0$. Top panels: fully observed lifetimes. Bottom panels: right censored and left truncation is included. Two different combinations of sample size and Gompertz parameters for generating observed lifetimes are used. Vertical red lines depict significance level selected for this study (5%) and the associated text presents the percentage of correctly selected Gompertz at that level.

Figure 4 is more explicit about the difficulty in discriminating between a Gompertz hazard and a mortality plateau. When lifetimes are not fully observed for censoring and truncation mechanisms, the odds of retrieving the true underlying Gompertz model are really low. For instance, if we take sample size and Gompertz parameters close to those presenting in Barbi et al. (2018b), $(N, a, b) = (4000, 0.7, 0.141)$, a likelihood-ratio test at 5% significance level would not reject the hypothesis $b = 0$ only 4.6% of the 500 simulated datasets.

Slight reverse trends in the percentages are visible in the lower parts of the panels in Figure 4. This is mainly due to the constant hazard for time to censoring which penalizes more datasets with smaller sample sizes. However, with a relatively higher underlying rate-of-aging, the number of censoring and truncated individuals decreases. Hence, trends in

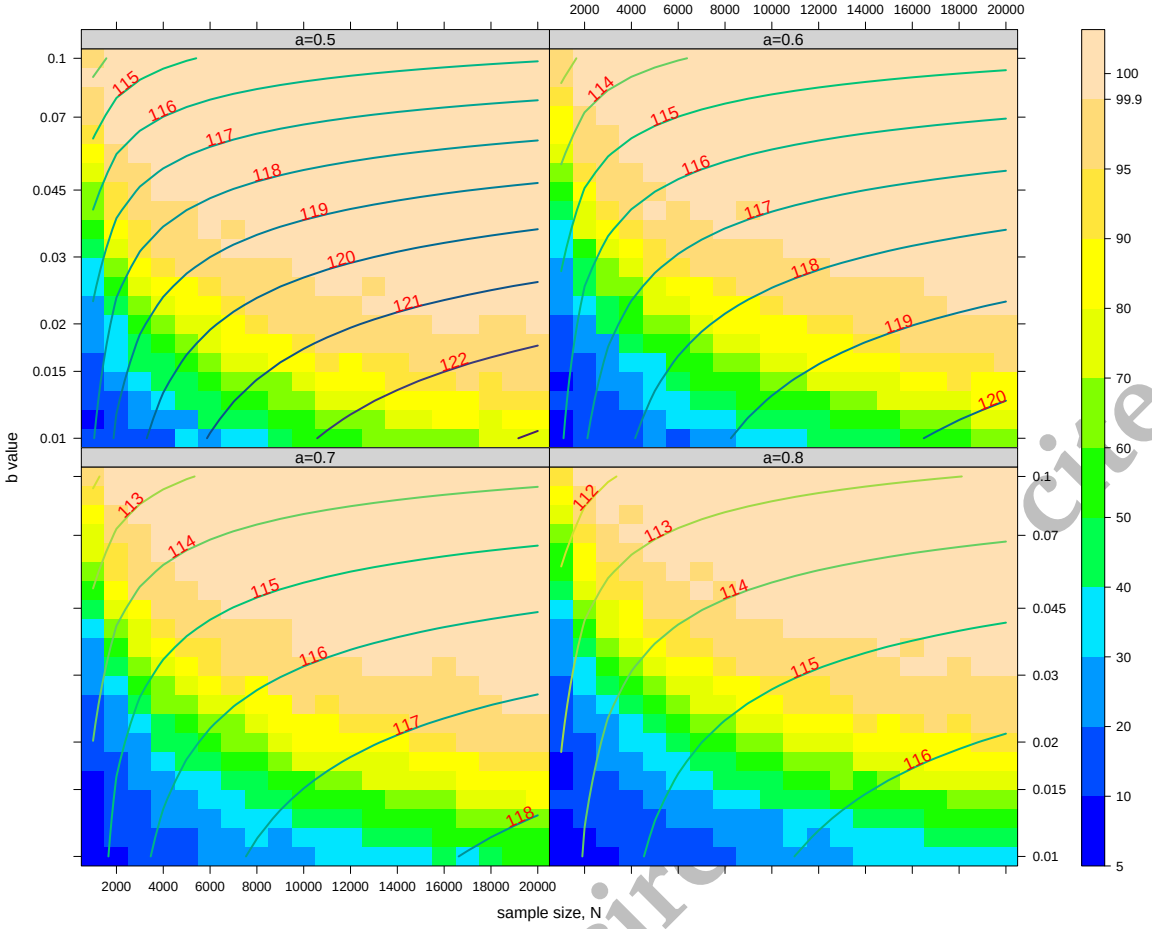

Figure 3: Percentage of correctly selected Gompertz by likelihood-ratio test with significance level 5% among 500 simulated datasets, by different sample size and true Gompertz parameters. Fully observed lifetimes. Contour lines of the associated theoretical highest attained age are superimposed.

proportion of correctly detected models become monotonically increasing after a certain level of b which depends upon sample size and starting level of mortality a.

In addition to the percentages of correctly selected model, Figures 3 and 4 present another layer of information: we included contour lines of the theoretical highest attained age, ω, for any combination of the triplet (N, a, b). Following Thatcher (1999), we consider ω as the extreme value of a sample of size N, drawn from a Gompertz distribution with known parameters a and b. Consequently, the highest attained age has a probability distribution and we portrayed in Figures 3 and 4 the isolines of the modal values of this distribution, i.e. the most probable highest age in which there is only one survivor among the starting N.

Theoretical highest attained age and knowledge about observed longevity records in history helps in dismissing possible simulated scenarios as implausible. For instance, by assuming Jeanne Calment as the oldest person ever with a lifespan of 122 years and 164 days, and by considering the 19,000 IDL cases as a complete database of people older than 105, it is unreasonable to admit a Gompertz distribution with starting level of mortality higher than 0.5 and with a rate-of-aging above 0.015.

A second illustration could be based on age 117 which has been outlived by the majority

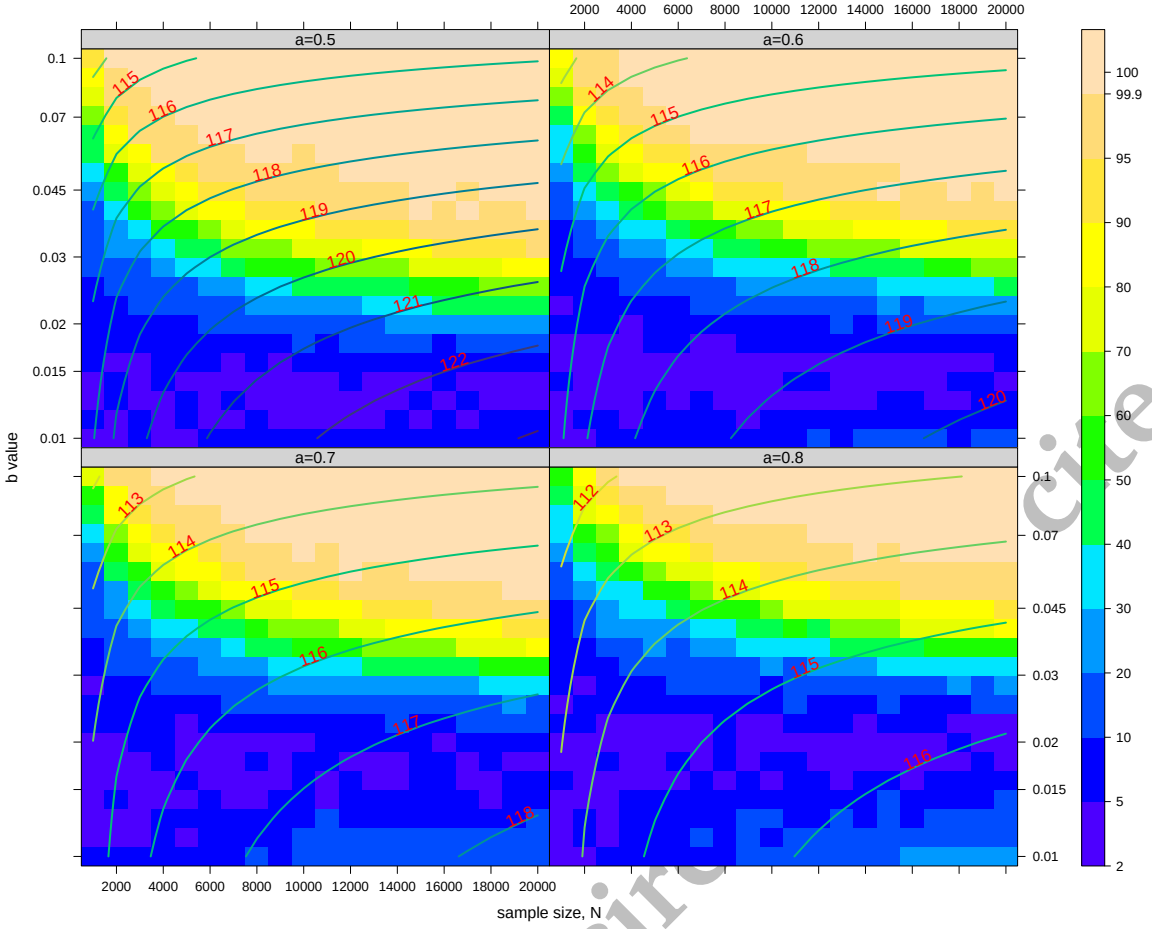

Figure 4: Percentage of correctly selected Gompertz by likelihood-ratio test with significance level 5% among 500 simulated datasets, by different sample size and true Gompertz parameters. Right censoring and left truncated observations. Contour lines of the associated theoretical highest attained age are superimposed.

of the top-ten verified oldest supercentenarians. In this case, a Gompertz distribution with $a = 0.8$ seems highly implausible. Similarly, scenarios with $a = 0.5$ and $N > 5,000$ are reasonable only when the rate-of-aging is below 0.06/0.07. If mortality at age 105 is 0.6 (0.7), we can admit as plausible only rate-of-aging below 0.04 (0.025). As expected, for a given a, the highest admissible b is quickly decreasing when the sample size gets smaller, e.g. for $a = 0.7$ and 20,000 observations above age 105, we could expect the actually observed supercentenarians when $b \leq 0.0251$, if N decreases to 8,000 a reasonable rate-of-aging should be smaller than 0.01.

Whereas both percentage of correctly selected Gompertz and highest attained age increases with N, these two phenomena behave differently with respect to the Gompertz parameters. Consequently, the (N, a, b)-combinations which are implausible by looking at the observed oldest people in the world corresponds to scenarios with high probability of selecting the underlying model. In other words, we may be able to clearly discriminate a Gompertz distribution from a mortality plateau only when the true Gompertz hazard is highly improbable, given the historical and verified longevity records.

For illustrative purposes Table 1 presents the percentages of datasets which will be selected

Table 1: Percentage of correctly selected Gompertz by likelihood-ratio test with significance level 5% among 500 simulated datasets from a Gompertz distribution with sample size $N = 19,000$, fully observed lifetimes and observations with right censoring and left truncation. The last four columns present the theoretical highest attained age associated to the Gompertz parameters with sample size $N = 19,000$. The sample size N is about the overall number of observations available in IDL.

	Fully obs lifetimes				Right Cens + Left Trunc				Highest attained age			
b \ a	0.5	0.6	0.7	0.8	0.5	0.6	0.7	0.8	0.5	0.6	0.7	0.8
0.0100	78.8	61.6	51.8	34.8	7.0	10.8	17.2	24.6	123.0	120.2	118.2	116.6
0.0112	82.6	70.6	55.4	47.6	5.4	9.4	15.4	14.4	122.8	120.1	118.1	116.5
0.0126	91.2	82.4	68.4	57.4	4.4	6.2	9.6	12.0	122.6	119.9	118.0	116.4
0.0141	95.6	88.6	75.0	66.8	4.2	5.2	5.8	10.0	122.4	119.8	117.8	116.4
0.0158	99.4	93.4	87.0	76.6	9.4	4.8	6.0	7.0	122.1	119.6	117.7	116.2
0.0178	99.8	97.4	93.4	80.6	15.6	8.8	5.0	6.2	121.9	119.4	117.6	116.1
0.0200	99.8	99.8	96.4	91.4	26.2	16.0	7.0	4.6	121.6	119.2	117.4	116.0
0.0224	100.0	100.0	99.0	96.6	44.4	22.0	11.6	5.0	121.3	119.0	117.2	115.9
0.0251	100.0	100.0	100.0	98.4	62.8	34.6	22.6	14.0	121.0	118.7	117.1	115.7
0.0282	100.0	100.0	100.0	99.8	82.4	61.2	37.4	18.6	120.7	118.5	116.9	115.6
0.0316	100.0	100.0	100.0	100.0	93.8	78.4	55.0	34.8	120.3	118.2	116.6	115.4
0.0355	100.0	100.0	100.0	100.0	98.4	90.2	78.2	54.6	119.9	117.9	116.4	115.2
0.0398	100.0	100.0	100.0	100.0	99.8	98.4	91.4	76.2	119.5	117.6	116.2	115.0
0.0447	100.0	100.0	100.0	100.0	100.0	99.8	91.4	76.2	119.1	117.3	115.9	114.8
0.0501	100.0	100.0	100.0	100.0	100.0	100.0	100.0	96.6	118.7	117.0	115.7	114.6
0.0562	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	118.3	116.6	115.4	114.4

correctly as generating by a Gompertz distribution for a sample size $N = 19,000$, as a function of a and b. We show this value of N since it is close to the total number of cases over 105 in the IDL. Again keep in mind that these percentages represents outcomes in either an ideal world (with fully observed lifetimes) or in relatively favorable conditions (with about 85% of the underlying lifetimes observed): much smaller values would be obtained with heavy truncation and censoring mechanisms, frequent in actual datasets. The last columns present the highest attained age expected for $N = 19,000$ and the corresponding Gompertz parameters.

3.2 Describing mortality age-pattern without functional assumptions

In this section, we will present the estimated hazards from a non-parametric approach by which no assumption is made on the functional form of the underlying force of mortality. Instead of simulating 500 times all (N, a, b)-triplets, we will present 3 case-studies. Given a parameter combination, we randomly select a dataset with both fully and partially observed simulated lifetimes. Equal conclusions might be drawn by selecting any of the alternative 500 simulated datasets within the same combination of parameters. As order of differences in the penalty term, we adopted both $d = 1$ and $d = 2$.

From the previous simulation setting, we chose instances with different percentages of correctly selected Gompertz model. Specifically, we take the following three combinations of (N, a, b): $(20000, 0.5, 0.0708)$, $(10000, 0.6, 0.0178)$ and $(3000, 0.8, 0.0112)$, corresponding to

100, 82.4 and 12.2% of datasets which are selected correctly as generating by a Gompertz distribution, when we assume to observe all lifetimes from 105 to death. When censoring and truncation mechanisms are included in the simulation settings, the same triplets lead to 100, 7.2 and 5% probability to correctly detect the true underlying Gompertz distribution. Moreover the highest attained ages for these sets of parameters are 117.4, 118.6 and 114.3 years.

Figure 5 presents the outcomes of this second simulation study. Left (right) panels present the true and smooth hazards when we assume fully (partially) observed lifetimes. From top to bottom panels, we have simulation setting from which we obtain a decreasing probability to correctly select a Gompertz distribution when it would be opposed to a mortality plateau. Along the hazards, gray bars identify distributions of deaths in each bin j. Right y-axes present the associated values. This additional information helps to better acknowledge the different simulation settings in terms of generated distributions as well as the expected rare observations in its right-tail. As consequence, in all instances, confidence intervals of the estimated smooth hazards tend to increase over age.

In estimating a smooth hazard by a penalized likelihood approach, choice about the order of differences in the penalty term is necessary (cf. Section 2.2). Whereas this decision is irrelevant when we model large sample size (Eilers and Marx, 2010), it become essential in our context. With few information from data, large smoothing parameter λ will be selected and consequently a plateau (Gompertz) mortality will be estimated when the penalty term is constructed with first (second) order of difference. This is an obvious outcome of any non-parametric approach: we search for an optimum between goodness-of-fit and parsimony in the estimated hazard. Hence, when few information is provided by the data, prior beliefs increase their importance in the final outcomes. In a penalized likelihood approach these beliefs are expressed by the order of difference in the penalty term.

Instead of selecting a certain order of difference, we present estimated smooth hazards in both cases. Figure 5 shows estimated smooth hazard for both $d = 1$ (in red) and $d = 2$ (in green). When we deal with a relatively large sample size (top panels), both procedures produce similar results up to about age 111. With fewer individuals the age in which estimated hazards from $d = 1$ and $d = 2$ diverge tend to decrease. However for all cases the 95% confidence intervals of the estimated hazards with $d = 1$ are mainly within the corresponding interval when we use $d = 2$, especially at the highest ages.

Confidence intervals allow us to gauge the ages in which a gompertzian hazard is distinguishable from a mortality plateau. Specifically, in top panels and in left-central panel, a Gompertz model with an increasing risk of dying seems identifiable before age 110, afterward and in all other settings and regardless the age, it will be impossible to differentiate between the two models, i.e. confidence intervals for estimated hazards with both first and second order of differences in the penalty term always contain a constant trajectory in the risk of dying.

At a glance we obtain seemingly oversmooth estimated trajectories above certain ages and/or when dealing with relatively small sample datasets. This is neither a consequence of the proposed penalized approach nor of the Akaike Information Criterion employed to select the amount of smoothness. On the contrary, the smoothing parameter selected by AIC tends to overfit our data. As example see the top panels of Figure 5 when sufficient information is provided by the data: smooth estimated hazards will pick up non-linear mortality patterns which are solely due to random fluctuations since data are simulated by a simpler Gompertz model.

When estimation is based on fully observed individuals (left panels), true underlying

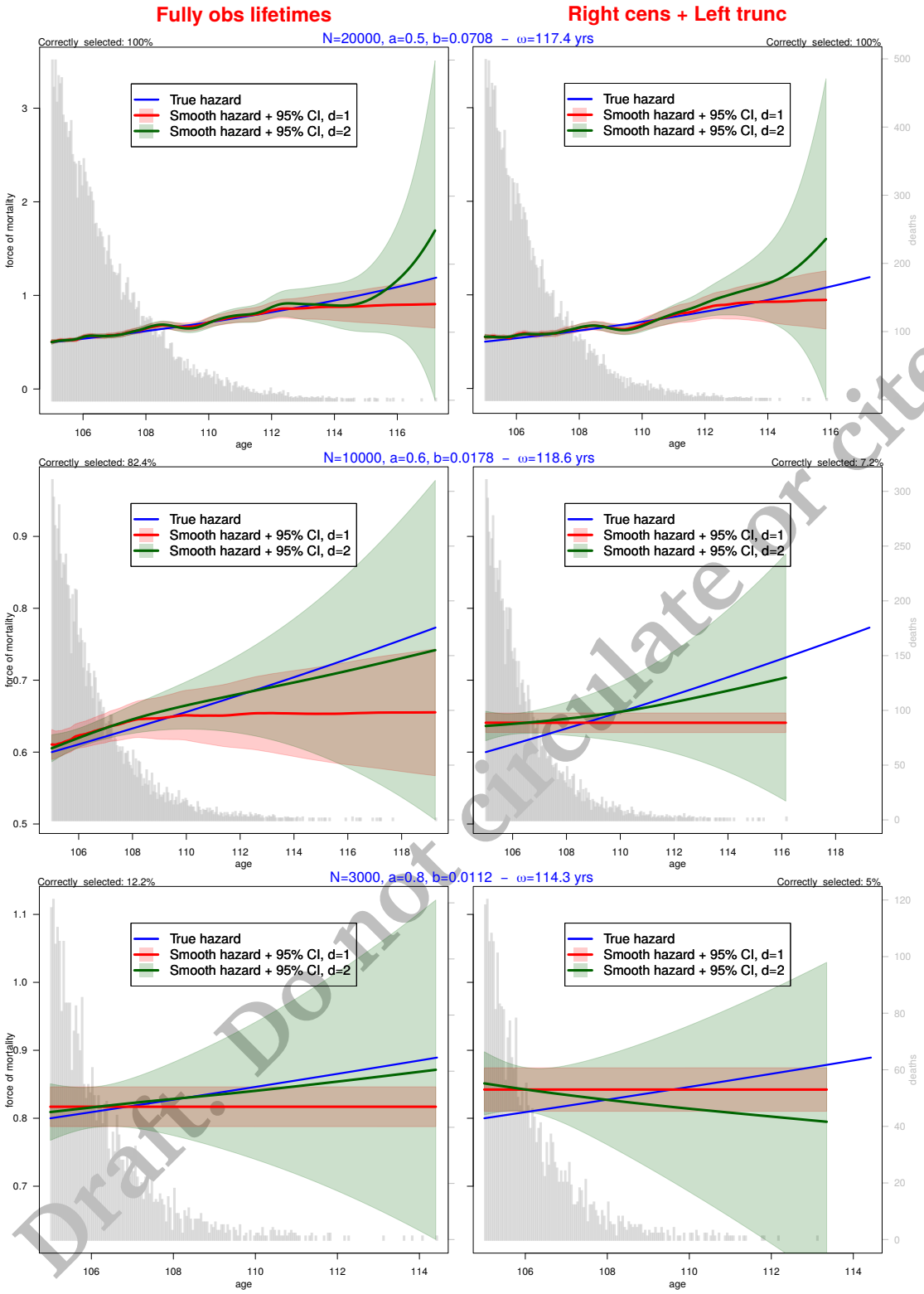

Figure 5: Non-parametric estimated hazard for three datasets with different sample size and true Gompertz parameters. Left (right) panels present simulation settings in which we assume fully (partially) observed lifetimes. Highest attained ages, ω, associated to each (N, a, b)-triplets as well as the percentages of correctly selected Gompertz from the previous simulation study are also displayed.

hazard (blue line) has higher chances to be include within the 95% confidence interval of the estimated hazards. Differences in mortality age-patterns estimated from either fully or partially observed lifetimes are mainly evident at the edge of the age ranges: left truncation and right censoring mechanisms lead to less accuracy in the estimation at early and later ages, respectively. See for instance how confidence intervals in the right panel tends to decrease just above age 105.

4 Conclusions

In recent years, the attempt to understand mortality at extreme ages have prompted a plethora of controversial studies. However, none of them have attempted to evaluate the statistical foundations of any analysis in which few individuals are available and experience exceptionally high risk of dying. In this paper we tackle the question from an alternative perspective: rather than providing an estimation of mortality age-pattern using actual data, we set up a rigorous simulation study. This approach allows us to disentangle all aspects behind a reliable estimation and it provides solid bases for evaluating when and what it can be accurately known when one faces the problem of mortality analysis at extreme ages.

Most of the current debate is about the existence of a mortality plateau after age 105 as opposite to a continuous increase, commonly described by a Gompertz model. We present two analyzes in which we generate individual lifetimes from a true underlying Gompertz distribution. In the first part we assess when we can really discriminate between a Gompertz hazard and a mortality plateau. As a second step, we free our estimated hazards from any functional form and we describe simulated data by a non-parametric approach. This allows us to comprehend up to which age we are able to clearly recognize a specific mortality pattern.

Different Gompertz parameters and sample sizes are considered to produce possible scenarios. We adopted a range of mortality patterns which largely includes estimated trajectories from previous studies and current populations. Meanwhile corresponding highest attained ages are juxtaposed to test plausibility of the simulated settings. Moreover, we perform our study on both fully and partially observed lifetimes, simulating eventual censoring and truncation times, too.

A model describing a mortality plateau is nested in a Gompertz hazard, therefore a conventional likelihood-ratio tests can be used to either accept or reject the additional Gompertz parameter. We performed this test on 500 simulated datasets for each parameters combination and for both fully and partially observed individual lifetimes. We clearly show that a relatively high percentage of correctly selected Gompertz model can be obtained only in combination of large sample sizes or high value of the true underlying Gompertz rate-of-aging. Whereas we could envisage and work to obtain larger datasets in the future, high rate-of-aging has never been observed in mortality data at very high ages, also considering the high starting level of mortality at age 105, i.e. death rates above 0.5.

A more discouraged outlook is obtained when lifetimes are assumed to be only partially observed, phenomenon very common in actual data. Specifically, we show that chance to identify a Gompertz instead of a mortality plateau when the true model is a Gompertz dramatically decreases when censoring and truncation mechanisms are included. Finally, most of the combinations of Gompertz parameters and sample sizes in which we could eventually have more than 50% probability to correctly identify a Gompertz correspond to implausible scenarios: the associated theoretical highest attained ages have been overcame by well-documented examples of oldest humans.

Moving from a dichotomous (plateau vs. Gompertz) to a broader framework, we generalized the Piecewise constant hazard model to estimate smooth mortality age-patterns. Free from any parametric structure, this approach allows us to characterize accurately the force of mortality and to monitor at which age a relatively complex pattern can be identified. Outcomes on relevant simulated case-studies show that only with the selected largest sample size (20,000) and up to age 110, we would be able to pinpoint a gompertzian trajectory. For higher ages and smaller sample size, confidence interval associated to the smooth hazard will always include a mortality plateau. Also in this setting, the presence of right censored and left truncated observations reduces the possibility to recognize mortality age-patterns more complex than a constant hazard.

In conclusion, we could rephrase Heisenberg's uncertainty principle for longevity studies: the more precisely the age-pattern of a population is determined, the less probable is its existence, and vice versa. Well, at least in our current world.

References

Antero-Jacquemin, J. S., G. Berthelot, A. Marck, P. Noirez, A. Latouche, and J. F. Toussaint (2015). Learning From Leaders: Life-span Trends in Olympians and Supercentenarians. *The journals of gerontology. Series A, Biological sciences and medical sciences 70*, 944–949.

Barbi, E., F. Lagona, M. Marsili, J. W. Vaupel, and K. W. Wachter (2018a). Response to Comment on "The plateau of human mortality: Demography of longevity pioneers". *Science 362*, eaav3229.

Barbi, E., F. Lagona, M. Marsili, J. W. Vaupel, and K. W. Wachter (2018b). The plateau of human mortality: Demography of longevity pioneers. *Science 360*(6396), 1459–1461.

Beltrán-Sánchez, H., S. N. Austad, and C. E. Finch (2018). Comment on "The plateau of human mortality: Demography of longevity pioneers". *Science 361*, eaav1200.

Brown, N. J. L., C. J. Albers, and S. J. Ritchie (2017). Contesting the evidence for limited human lifespan. *Nature 546*, E6–E7.

Camarda, C. G., L. H. K. Dang, F. Meslé, J.-M. Robine, and J. Vallin (2018). Premature claim of a plateau of human mortality: the role of sample size. *Science*, eLetter. (5 October 2018).

Dang, L., C. G. Camarda, F. Meslé, N. Ouellette, J.-M. Robine, and J. Vallin (2019). Survival Analysis on Mortality Data at Oldest Ages: First Results on Longevity Pioneers in France. In *Population Association of America 2019 Annual Meeting*, 10-13 April, Austin, TX.

de Beer, J., A. Bardoutsos, and F. Janssen (2017). Maximum human lifespan may increase to 125 years. *Nature 546*, E16–E17.

Devroye, L. (1986). *Non-Uniform Random Variate Generation*. New York: Springer-Verlang.

Dong, X., B. Milholland, and J. Vijg (2016). Evidence for a limit to human lifespan. *Nature 538*, 257–259.

Eilers, P. H. C. and J. J. Goeman (2004). Enhancing scatterplots with smoothed densities. *Bioinformatics 20*(5), 623–628.

Eilers, P. H. C. and B. D. Marx (1996). Flexible Smoothing with B-splines and Penalties (with discussion). *Statistical Science 11*, 89–102.

Eilers, P. H. C. and B. D. Marx (2010). Splines, Knots, and Penalties. *Wiley Interdisciplinary Reviews: Computational Statistics 2*, 637–653.

Einmahl, J. J., J. H. J. Einmahl, and L. de Haan (2019). Limits to Human Life Span Through Extreme Value Theory. *Journal of American Statistical Association 114*, 1075–1080.

Gampe, J. (2010). Human mortality beyond age 110. In H. Maier, J. Gampe, B. Jeune, J.-M. Robine, and J. W. Vaupel (Eds.), *Supercentenarians*. Berlin: Springer-Verlag.

Gavrilov, L. A. and N. S. Gavrilova (2019). Late-life mortality is underestimated because of data errors. *PLOS Biology 17*, e3000148.

Gavrilov, L. A., V. N. Krut'ko, and N. S. Gavrilova (2017). The future of human longevity. *Gerontology 63*, 524–526.

Gavrilova, N. and L. A. Gavrilov (2011). Mortality Measurement at Advanced Ages: A Study of the Social Security Administration Death Master File. *North American Actuarial Journal 15*, 432–447.

Gavrilova, N. and L. A. Gavrilov (2015). Biodemography of Old-Age Mortality in Humans and Rodents. *Journals of Gerontology Series A - Biological Sciences and Medical Sciences 70*, 1–9.

Gavrilova, N. S., L. A. Gavrilov, and V. N. Krut'ko (2017). Mortality Trajectories at Exceptionally High Ages: A Study of Supercentenarians. In *2017 Living to 100 Monograph*, pp. 1–17. Society of Actuaries.

Gompertz, B. (1825). *On the nature of the function expressive of the law of human mortality.* 115: 513-585. London, UK: Philosophical Transactions Royal Society.

Holford, T. R. (1980). The analysis of rates and survivorship using log-linear models. *Biometrics 36*, 299–305.

Hughes, B. G. and S. Hekimi (2017). Many possible maximum lifespan trajectories. *Nature 546*, E8–E9.

Human Mortality Database (2019). *University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany).* Available at `www.mortality.org`. Data downloaded on July 2019.

International Database on Longevity (2019). *Max Planck Institute for Demographic Research (Germany), and Institut national d'études démographiques (France).* Available at `www.supercentenarians.org`. Data downloaded on April 2019.

Jackson, C. (2016). flexsurv: A platform for parametric survival modeling in R. *Journal of Statistical Software 70*, 1–33.

Keiding, N. (2018). Comments to Rootzén & Zholud: Human life is unlimited - but short. *Extremes 21*, 383–386.

Klein, J. P. and M. L. Moeschberger (2003). *Survival Analysis. Techniques for Censored and Truncated Data* (2nd ed.). New York: Springer-Verlag.

Krivobokova, T., C. Crainiceanu, and G. Kauermann (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics 17*(1), 1–20.

Lenart, A., J. M. Aburto, A. Stockmarr, and J. W. Vaupel (2018). The human longevity record may hold for decades. (11 September 2018).

Lenart, A. and J. W. Vaupel (2017). Questionable evidence for a limit to human lifespan. *Nature 546*, E13–E14.

Medford, A. (2018). Strong evidence for a plateau: but where is it? *Science*, eLetters. (31 August 2018).

Milholland, B., X. Dong, and J. Vijg (2017). "best-guess" MRAD Provides Robust Evidence for a Limit to Human Lifespan: Reply to de Grey (Rejuvenation Res. 2017;20:261262). *Rejuvenation Research*.

Milholland, B., X. Dong, and J. Vijg (2018). Evidence for limited lifespan in semisupercentenarians. *Science*, eLetter. (10 October 2018).

Newman, S. J. (2018a). Errors as a primary cause of late-life mortality deceleration and plateaus. *PLOS Biology 16*, e2006776.

Newman, S. J. (2018b). Plane inclinations: A critique of hypothesis and model choice in Barbi et al. *PLOS Biology 16*, e3000048.

Olshansky, S. J. and B. A. Carnes (2018). Mortality plateaus and unlimited lifespan: let's get real. *Science*, eLetters. (17 July 2018).

Rau, R., M. Ebeling, F. Peters, C. Bohk-Ewald, and T. I. Missov (2017). Where is the level of the mortality plateau? In *2017 Living to 100 Monograph*, pp. 1–22. Society of Actuaries.

Rootzén, H. and D. Zholud (2017). Human life is unlimited - but short. *Extremes 20*, 713–728.

Rozing, M. P., T. B. L. Kirkwood, and R. G. J. Westendorp (2017). Is there evidence for a limit to human lifespan? *Nature 546*, E11.

Thatcher, A. R. (1999). The long-term pattern of adult mortality and the highest attained age (with discussion). *Journal of Royal Statistical Society. Series A 127*, 5–43.

Vijg, J. and E. Le Bourg (2017). Aging and the Inevitable Limit to Human Life Span. *Gerontology 63*, 432–434.

Wachter, K. (2018). Hypothetical Errors and Plateaus: A Response to Newman. *PLoS Biology 16*, e3000076.

Witten, M. and W. Satzer (1992). Gompertz model survival parameters: Estimation and sensitivity. *Applied Mathematics Letters 5*, 7–12.

Wood, S. N. (2006). *Generalized Additive Models. An Introduction with R*. Chapman & Hall.