8-31-2020

Modeling Age Patterns of Under-5 Mortality: Results From a Log-Quadratic Model Applied to High-Quality Vital Registration Data

Michel Guillot
University of Pennsylvania, miguillo@upenn.edu

Julio Romero Prieto
London School of Hygiene and Tropical Medicine, julio.romero-prieto@lshtm.ac.uk

Andrea Verhulst
University of Pennsylvania, verhulst@sas.upenn.edu

Patrick Gerland
United Nations, gerland@un.org

Recommended Citation

Guillot, Michel, Julio Romero Prieto, Andrea Verhulst, and Patrick Gerland. 2020. "Modeling Age Patterns of Under-5 Mortality: Results From a Log-Quadratic Model Applied to High-Quality Vital Registration Data." *University of Pennsylvania Population Center Working Paper (PSC/PARC)*, 2020-54. https://repository.upenn.edu/psc_publications/54.

Modeling Age Patterns of Under-5 Mortality: Results From a Log-Quadratic Model Applied to High-Quality Vital Registration Data

Abstract

Information about how the risk of death varies with age within the 0-5 age range represents critical evidence for guiding health policy. This paper proposes a new model for summarizing regularities about how under-5 mortality is distributed by detailed age. The model is based on a newly compiled database that contains under-5 mortality information by detailed age in countries with high-quality vital registration systems, covering a wide array of mortality levels and patterns. The model uses a log-quadratic approach, predicting a full mortality schedule between age 0 and 5 on the basis of only 1 or 2 parameters. With its larger number of age groups, the proposed model offers greater flexibility than existing models both in terms of entry parameters and model outcomes. We present applications of this model for evaluating and correcting under-5 mortality information by detailed age in countries with problematic mortality data.

Keywords

under-5 mortality, neonatal mortality, model life tables, mortality models, age patterns of mortality, indirect methods

Disciplines

Demography, Population, and Ecology | Family, Life Course, and Society | Inequality and Stratification | Social and Behavioral Sciences | Sociology

Modeling Age Patterns of Under-5 Mortality:
Results From a Log-Quadratic Model Applied to High-Quality Vital Registration Data

Michel Guillot
University of Pennsylvania &
French Institute for Demographic Studies (INED)

Julio Romero Prieto
London School of Hygiene and Tropical Medicine

Andrea Verhulst
University of Pennsylvania

Patrick Gerland
United Nations

Corresponding author:
Michel Guillot, Population Studies Center, University of Pennsylvania, 239 McNeil Building, 3718 Locust Walk, Philadelphia PA 19104.
Email: miguillo@sas.upenn.edu. Tel: 215-573-3655. Fax: 215-898-2124.

Acknowledgements: Research reported in this manuscript was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) of the National Institutes of Health (NIH) under award number R01HD090082. Additional funding was provided by the UNICEF. Earlier versions of this paper were presented at the 2019 PAA meetings in Austin, TX, the Human Mortality Database Symposium in Berlin, the London School of Hygiene and Tropical Medicine, and the French Institute for Demographic Studies. The authors would like to thank participants of these seminars, as well as Ken Hill, Sam Preston, Georges Reniers, Elizabeth Wrigley-Field, and members of the Inter-Agency Group for Child Mortality Estimation (IGME) for their feedback. We also thank Sharon Chan for her contribution to the data collection, as well as the National Statistical Office of Jordan and UN-ESCWA for providing the vital registration data for Jordan. The views expressed in this paper are those of the authors and do not necessarily reflect the views of the United Nations.

Modeling Age Patterns of Under-5 Mortality:
Results From a Log-Quadratic Model Applied to High-Quality Vital Registration Data

Abstract

Information about how the risk of death varies with age within the 0-5 age range represents critical evidence for guiding health policy. This paper proposes a new model for summarizing regularities about how under-5 mortality is distributed by detailed age. The model is based on a newly compiled database that contains under-5 mortality information by detailed age in countries with high-quality vital registration systems, covering a wide array of mortality levels and patterns. The model uses a log-quadratic approach, predicting a full mortality schedule between age 0 and 5 on the basis of only 1 or 2 parameters. With its larger number of age groups, the proposed model offers greater flexibility than existing models both in terms of entry parameters and model outcomes. We present applications of this model for evaluating and correcting under-5 mortality information by detailed age in countries with problematic mortality data.

Keywords: under-5 mortality, neonatal mortality, model life tables, mortality models, age patterns of mortality, indirect methods.

Introduction

The Under-5 Mortality Rate (U5MR) is a key and widely used indicator of child health (United Nations 2011; United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) 2019b; Wang et al. 2016; You et al. 2015), but it conceals important information about how this mortality is distributed by age from birth up to the fifth birthday (Guillot et al. 2012; Hill 1995; Mejía-Guevara et al. 2019). For better understanding and monitoring of child health, it is critical to examine how the risk of death varies within the first five years of life. This includes age breakdowns beyond the standard cut-off points of 28 days (for neonatal mortality) and 1 year (for infant mortality). In many populations, however, the age pattern of under-5 mortality is not well known. Low- and middle-income countries, in particular, lack the high-quality detailed vital registration information necessary for the analysis of such age patterns (Mikkelsen et al. 2015). Sample surveys collecting retrospective birth histories, such as Demographic and Health Surveys (DHS), do not satisfactorily fill this gap, because they are subject to potential biases that are particularly consequential for estimating age patterns (Hill 1995; Lawn et al. 2008). This makes the need for high-quality information on age patterns of under-5 mortality even more critical, since regularities in these age patterns can be used as a powerful tool for evaluating and correcting estimates when data are deficient.

The goal of this paper is to propose a new model for summarizing regularities about how under-5 mortality is distributed by detailed age in human populations. This model is based on the Under-5 Mortality Database (U5MD), a newly compiled database that contains under-5 mortality information by detailed age in countries with high-quality vital registration systems, covering a wide array of mortality levels and patterns. Building on previous work by Wilmoth et al. (2012), this model uses a log-quadratic approach, predicting a full mortality schedule between age 0 and 5 on the basis of only 1 or 2 parameters. We present applications of this model for evaluating and correcting under-5 mortality information by detailed age in countries with deficient mortality data.

This paper builds on the model life tables literature. Model life tables summarize regularities in how mortality varies by age in human populations. They represent a useful framework for our purpose because they allow the estimation of arrays of age-specific mortality rates or probabilities on the basis of only one or two mortality indicators, chosen as entry parameters (United Nations 1988). Two sets of model life tables are considered classic in the field: one set was developed by Coale and Demeny (Coale & Demeny 1966; Coale et al. 1983) and the other by the United Nations Population Division (1982). These two sets are still commonly used today, including for estimating the infant mortality rate (IMR) on the basis of U5MR (United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) 2019a). Current usage of the Coale and Demeny and the United Nations model life tables for estimating patterns of under-5 mortality, however, is affected by several important drawbacks. First, these model life tables only offer 0 vs. 1-4 as an age breakdown for under-5 mortality. This is insufficient for most purposes, including for the estimation of neonatal mortality or mortality in non-standard age ranges. (One model that contains additional age details is Bourgeois-Pichat's "biometric" model (Bourgeois-Pichat 1951). This model, however, focuses on the first 12 months of age only and has been shown to poorly fit data in a variety of contexts (Galley & Woods 1998; Knodel & Kintner 1977; Lantoine & Pressat 1984; Lynch et al. 1998; Manfredini 2004).) Second, existing model life tables rely on rather old data, with the most recent information dating back to the early 1980s. Third, these model life tables summarize age patterns as "families," based on regional groupings, and thus have a discrete rather than continuous nature. More recent developments in the model life tables literature include Murrays et al.'s (2003) modified logit system, Wilmoth et al.'s (2012) log-quadratic model, and Clark's (2019) SVD-component model. These models improve on many of the weaknesses of the classic model life tables, including the use of a continuous rather than discrete parameter for describing variations in mortality shapes, and the use of more recent data for deriving model coefficients. However, Murrays et al.'s (2003) and Wilmoth et al.'s (2012) models are still constrained by the 0 vs. 1-4 age breakdown for the under-5 age range, and Clark's (2019) model does not provide details below single-year age groups. Our paper extends existing model life tables by: (1) using a newly compiled database that has greater age detail than the ones on which existing model life tables were derived; (2) explicitly expanding the number of age groups in the model, especially in the first year of life, allowing more flexibility than existing models both in terms of entry parameters and model outcomes. Our model offers a number of applications that are not feasible with existing model life tables, including the possibility of detecting and adjusting for underestimation of neonatal mortality.

A New Database for Under-5 Mortality by Detailed Age

Description of the Database

The model proposed in this paper is based on the Under-5 Mortality Database (U5MD), a newly compiled database for under-5 mortality by detailed age drawn from high-quality Vital Registration (VR) data. This database contains 1,652 annual distributions of under-5 deaths by sex and detailed age (days, weeks, months, trimesters, and years), representing 25 countries over a time window spreading from the second half of the nineteenth century to recent years (1841-

2016). The list of available country-years is provided in Table 1. This section summarizes how this database was built and harmonized. Full details are available in the Supplementary Materials.

-- Table 1 about here --

Age distributions of deaths were obtained from two primary sources: (i) For historical periods (prior to 1970), these distributions were collected manually from archival sources such as national statistical yearbooks; and (ii) for periods from 1970 onwards, they were obtained electronically from a data repository compiled by the United Nations Statistical Division.

Country-years were selected based on two criteria: (1) the quality of data; and (2) the availability of detailed age breakdowns. For the data quality criterion, the U5MD used the criterion of virtual completeness of death registration and census data determined by the *Human Mortality Database* (HMD) (Barbieri et al. 2015). This means that we only considered country-years available in the HMD for inclusion in the U5MD. The HMD comprises mostly European countries (31) but also some other industrialized countries (9). However, we did not include all HMD countries in the U5MD. As discussed in the Supplementary Materials, we excluded countries of the former Eastern bloc due to well-documented concerns about the quality of the mortality data at early ages. Greece was also excluded for similar reasons (Agorastakis et al. 2017). In addition, Iceland and Luxembourg were removed due to the small size of the population leading to many zero cell counts in the narrow age group we focus on in this paper.

Regarding the detail of the age information, the minimum criteria for inclusion in the U5MD was the breakdown of infant deaths in terms of neonatal deaths (<1 month) vs. post-neonatal deaths

(1-11 months). The death distributions we collected typically included much finer age granularity, but the format of age intervals varied greatly across the primary sources of information. Deaths were tabulated unevenly by days, weeks, months, trimester, semester and years, and distributed over different age spans (first year of age only vs. larger age ranges up to the full first five years). In order to address this unevenness, we harmonized age groups into 22 age intervals with the following exact-age cut-off points: 0, 7, 14, 21, 28 days; 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 18, 21 months; 2, 3, 4, 5 years. The harmonization was carried out by interpolating cumulative age distributions of deaths using a method developed by Steffen (1990). We excluded 99 country-years at that stage due to insufficient age details during the first month for performing this interpolation (see Table 1 and the Supplementary Materials for details).

Our database was complemented by two pieces of information obtained directly from the HMD for the country-years covered in the U5MD: (1) raw death counts between exact ages 1 and 5, which we used to fill potential missing information in our database in that age range; (2) exposures to the risk of dying in person-years, by calendar year and by single year of age, calculated by the HMD from census and birth data (Wilmoth et al. 2017).

Age-specific deaths rates (${}_nM_x$) and corresponding probabilities of dying from birth to age x ($q(x)$) were computed for each of the 22 harmonized age intervals. Death rates were computed by dividing deaths by the exposure to the risk of death for each age interval and year. Since exposure terms were not available for age groups smaller than one year, we assumed a uniform distribution of exposure within each single-year age group. With this assumption, exposure terms are proportional to the length of the age interval *n* within each single-year age group. We then calculated cumulative probabilities of dying *q(x)* with the assumption that mortality rates were constant within each age interval, in which case $q(x+n) = 1-(1-q(x)) \cdot e^{-n \cdot {}_nM_x}$. This assumption is not very consequential given the small width of our age intervals.

Evaluation of the Quality of the U5MD

As discussed above, the U5MD includes a subset of country-years covered in the HMD, a source representing the gold standard in terms of VR mortality information. Nonetheless, when focusing on under-5 mortality by detailed age, questions remain about the quality of the reported information, especially for earlier periods (19th century and early 20th century) and for the neonatal age range. Neonatal deaths are known to be subject to underreporting, especially when they occur very soon after birth. This is due in part to ambiguities about what constitutes a live birth vs. a stillbirth. Discussions of international standards for defining live births vs. stillbirths started only in the 1920s under the impulse of the League of Nations (United Nations Department of Social Affairs - Population Division 1954), and distinguishing between live births and stillbirths remains a complex issue even today (Gourbin & Masuy-Stroobant 1995; Hug et al. 2019). This raises questions about how correctly this distinction was made during the earlier years covered in our database. Another source of underreporting arises from the fact that when a child death occurs before the recording of the corresponding live birth, the incentive to report these two events in civil registers is low. This further questions the quality of the reporting of neonatal deaths during the earlier periods of the database, at a time when most deliveries occurred at home (United Nations Statistical Office 1955).

Due to these data quality concerns, we performed an evaluation of the quality of the U5MD prior to estimating our model. Specifically, we performed plausibility checks, focusing on mortality during the neonatal period. We examined the relationship between age-specific mortality rates for the first, second, third and fourth week of life (${}_7M_{0(d)}$, ${}_7M_{7(d)}$, ${}_7M_{14(d)}$ and ${}_7M_{21(d)}$, respectively, with the letter *d* indicating that age is expressed in days) vs. the probability that a 28-day child will die prior to reaching age 5 years (q(28d,5y)), i.e., a mortality indicator not affected by mortality rates for the neonatal period. These relationships are shown in Figure 1.

--- Figure 1 about here ---

Figure 1 shows that for weeks 2, 3 and 4, there is a clear positive – almost log-linear – relationship between each weekly mortality rate and mortality between 28 days and 5 years. There is no large change in slope at any point in the relationship, including when q(28d,5y) is high, i.e., during the earlier years of our database. The mortality rate for the first week, however, has a drastically different relationship with q(28d,5y). While the relationship starts with a clear upward slope, there appears to be a flattening of the relationship as q(28d,5y) reaches high levels. For some individual country trajectories, we even found reversals in the relationship, depicting situations where decreases over time in reported mortality between 28 days and 5 years coincide with *increases* in the reported mortality rate for the first week.

These flattenings and reversals are suspicious for a number of reasons. First, the changes in slope take place during the earlier years in our database, with turning points typically occurring between WWI and WWII. These earlier years are the years for which the sources of errors are most likely to apply. Second, the changes in slope occur only for the first week, which is the week that is most subject to the sources of errors mentioned earlier. Weeks 2-4, which are less subject to these errors, show no such flattenings. Third, within the first week, changes in slope are most pronounced during days 0-3, which are the days most subject to errors (results not shown). The relationships are more log-linear for days 4-6, which are less subject to errors. Fourth, reversals and flattenings do not occur everywhere, suggesting that monotonic relationships between mortality for the first week (${}_{7}M_{0(d)}$) and mortality between ages 28d and 5y are biologically possible. In Switzerland, for example, the level of ${}_{7}M_{0(d)}$ keeps increasing together with q(28d,5y) as we go further back in time, with no signs of decrease in slope.

Taken altogether, these issues raise serious doubts about the quality of the early neonatal mortality data during the earlier years covered in the database. Rather than excluding all the data points above a given mortality level, we decided to take an intermediate approach that excludes long-lasting reversals in the ${}_{7}M_{0(d)}$ vs q(28d,5y) relationship. Specifically, we decided to remove, for a given country, all points to the right of a given point in the ${}_{7}M_{0(d)}$ vs q(28d,5y) scatterplot when there are 15 temporally consecutive points that are all below that one given point. This approach removes the most suspicious patterns while keeping the possibility of a decrease in slope at higher levels of q(28d,5y).

This exclusion criteria removes 318 country-years, shown in Table 1. As expected, the excluded country-years pertain mostly to the early years covered by the database: 19th century and early 20th century.

Final Database for Modeling Purposes

The final U5MD that we use for our model includes 1,235 country-years, by sex and for both sexes combined. These country-years cover a wide range of time periods and levels of under-5 mortality, from the late nineteenth century until today, with levels ranging from above 200 to less than 5 per 1000. A summary of the available country-years is available in Table 1 (last column), and full details are provided in Appendix Table A1.

Log-Quadratic Model for Age-Specific Mortality by Detailed Age between 0 and 5

Model Description

We propose a model able to predict a full mortality schedule by detailed age between 0 and 5 years with only two parameters, one depicting the overall *level* of under-5 mortality and the other depicting the *shape* of the age pattern of mortality within the 0-5 age range. This model is adapted from Wilmoth et al.'s (2012) log-quadratic model; it is based on the observation of log-quadratic relationships between the cumulative probability of dying from birth to age x, q(x), and the under-5 mortality rate, q(5y), for each detailed age x within the under-5 age range:

$$\ln[q(x)] = \mathrm{a}_x + \mathrm{b}_x \cdot \ln[q(5y)] + \mathrm{c}_x \cdot \ln[q(5y)]^2 + \mathrm{v}_x \cdot k \qquad (1)$$

As shown in Equation (1), the model includes a set of age-specific coefficients $\{\mathrm{a}_x, \mathrm{b}_x, \mathrm{c}_x, \mathrm{v}_x\}$, whose estimation we describe below. When $k = 0$, the model predicts a general pattern which is the average mortality schedule of the set of country-years included in the final U5MD. When k $\neq$ 0, the model adjusts the probabilities of dying in response to specificities in the age pattern of q(x) at a given level of q(5y), bearing in mind that $q(x)$ is a non-decreasing function of age. For a given level of $q(5y)$, depending on the value of k, the age pattern of mortality will be either "early", with relatively high levels of neonatal and infant mortality; or "late", when these levels are relatively low.

Note that unlike the Wilmoth et al. (2012) approach, our model involves cumulative probabilities of dying, q(x), rather than age-specific mortality rates, ${}_{n}M_{x}$, in the left-hand side of Equation (1). There are four advantages in doing so: (1) the predicted set of q(x) and its corresponding values of ${}_{n}M_{x}$ will always agree with the level of q(5y) that is chosen as predictor in the right-hand side of Equation (1); (2) the model will be more parsimonious, with 21 coefficients vs. 22 when using mortality rates; (3) the model will be less sensitive to fluctuations in the mortality schedule that could arise from misreported ages at death; (4) the model will directly predict classic mortality indicators such as early-neonatal, neonatal and infant mortality rates, which are in fact cumulative probabilities of dying (q(7d), q(28d) and q(12m), respectively). There is however one drawback in using cumulative probabilities of dying in this model: data errors at early ages such as underreporting of neonatal deaths will carry through the entire q(x) curve. This makes our rather conservative approach with respect to the inclusion of country-years in the final U5MD all the more important. Although our model predicts cumulative probabilities of dying rather than age-specific mortality rates, corresponding mortality rates can be easily recovered from the predicted q(x) values using the assumption of a constant force of mortality within each of our 22 small age intervals:

$${}_{n}M_{x} = -\frac{ln\left[\frac{1-q(x+n)}{1-q(x)}\right]}{n}.$$

While developing our model, we also explored the possibility of building a model based on Clark's (2019) more general SVD-component model. One of the main differences between the log-quadratic model and the SVD-component model is that the latter does not include a parametric assumption relating age-specific mortality to a mortality indicator like q(5y) chosen as the main explanatory variable. Instead, the SVD-component model is a linear sum of independent, age-varying vectors, like in a Principal Component Analysis (PCA) decomposition. After exploring both approaches, we decided to follow the log-quadratic approach because the parametric assumption was appropriate for the narrower (0 to 5) age range that is the focus here. This parametric assumption makes the log-quadratic model more parsimonious and easier to use when focusing on this younger age range.

Estimating the Coefficients $\{\mathrm{a_x}, \mathrm{b_x}, \mathrm{c_x}, \mathrm{v_x}\}$

The model coefficients in Equation (1) were estimated in two steps. The first step involved the estimation for each age x of the set of age-specific coefficients $\{\mathrm{a}_x, \mathrm{b}_x, \mathrm{c}_x\}$ regressing $q(x)$ against $q(5y)$ with OLS. This is shown in equation (2), with the subscript i indicating each country-year in our sample of $N = 1{,}235$ observations.

$$\ln[q_i(x)] = \mathrm{a}_x + \mathrm{b}_x \cdot \ln[q_i(5y)] + \mathrm{c}_x \cdot \ln[q_i(5y)]^2 + e_i(x) \qquad (2)$$

The second step uses the age-covariance of the residuals $e(x)$ in Equation (2) which informs about systematic deviations from the general pattern of mortality, for estimating the set of coefficients v_x. For this purpose, we estimated the covariance matrix of the residuals Ψ, whose element (z, y) is given by: $\Psi_{zy} = \frac{1}{N-3} \cdot \sum_{i=1}^{N} e_i(z) \cdot e_i(y)$. Following a common approach in demographic estimation (Clark 2019; Lee & Carter 1992; Wilmoth et al. 2012; Wilmoth 1990), we estimated the set of coefficients v_x as the first-orthonormal eigenvector (of V) resulting from a Singular Value Decomposition (SVD) applied to the covariance matrix: $\Psi = \mathrm{V} \cdot \Sigma \cdot \mathrm{U}$. The SVD provides a least squares solution to the principal components of the residuals, hence the first vector will account for the higher proportion of the overall covariance. In our case, the first eigenvalue accounts for the 87% of the total sum of eigenvalues.

Model Results

Table 2 shows the model coefficients for males, females, and both sexes estimated using the final U5MD. This table shows that as age x increases, b_x approaches 1 and c_x approaches zero. This is expected given that as x increases, q(x) approaches q(5y). At younger ages, however, we find significantly negative values of c_x. This reflects decreasing slopes in the relationship between q(x) and q(5y) at high levels of q(5y). Values of v_x all have the same negative signs. This is due to the fact that when an age pattern of mortality is late or early relative to the average, the entire q(x) curve is shifted up or down. The comparison of male vs. female coefficients shows that while values of c_x and v_x are very similar for each sex, values of a_x and b_x present sizeable differences, with male coefficients being systematically higher than the female ones. This means that at a given level of q(5y) and k, the model will produce and earlier age pattern of mortality for males.

--- Table 2 about here ---

These features of the model results are illustrated in Figure 2, which shows observed vs predicted values of q(7d), q(28d) and q(12m) when k=0 and when k=+/-1. Note that almost all data points used for estimating the model are included within this range of values for k.

--- Figure 2 about here ---

The model results are further illustrated in Figure 3, which shows how predicted values of q(x) (Panel A) and corresponding values of ${}_nM_x$ (Panel B) vary in response to changes in the level of q(5y) at a given level of k (=0 in this example). As the level of q(5y) changes from 100 to 10 per 1000, an increasing portion of under-5 mortality takes place below one year and below 28 days. This is a well-known regularity that reflects the transition from a situation with a high prevalence of infectious ("exogenous") causes of death that have an older age pattern to one in which infectious diseases have been virtually eliminated and the only remaining causes are congenital anomalies and perinatal conditions, i.e., "endogenous" causes that have a younger age pattern (Drevenstedt et al. 2008; Galley & Woods 1999; Liu et al. 2012; Rao et al. 2011). Examining the shape of the mortality curves in the right panel, we see that our model produces mortality patterns that monotonically decrease with age. This also reflects the regularities present in our database. Indeed the country-years included in the database do not present any systematic age-specific mortality reversals. As the level of q(5y) decreases, the entire mortality curve between 0 and 5 shifts down, with larger relative declines at older vs. younger ages.

--- Figure 3 about here ---

Figure 4, Panel A shows the effect of varying k on the q(x) curve at a given level of q(5y) (=100 per 1000 in this example). When k=+1, the entire q(x) curve is shifted down. This produces a "late" pattern of under-5 mortality, with lower levels of neonatal and infant mortality while q(5y) remains unchanged. Conversely, when k=-1, this produces an "early" pattern of under-5 mortality, with higher levels of neonatal and infant mortality.

--- Figure 4 about here ---

Figure 4 also shows corresponding effects of changing k on ${}_{n}M_{x}$ values between 0 and 5 (Panel B), with a zoom on the first 3 months (Panel C). The mortality curves in this figure all produce the same level of under-5 mortality (100 per 1000 in this example). Higher levels of mortality at some ages will thus necessarily have to be compensated by lower levels of mortality at some other ages. The resulting mortality crossover is visible in the right panel of Figure 4, which shows that the "tilting" age occurs during the second month of life. This implies that at this level of q(5y), the shape of the age pattern of mortality is entirely explained by the contrast between q(28d) vs. q(28d,5y). The age at which this crossover occurs in our model is however not constant but related to the level of under-5 mortality. The lower the level of q(5y), the earlier the crossing age. When q(5y) reaches a level around 50 per 1000, the crossover occurs during the second week, its lower limit. This means that at these lower levels of q(5y), the shape of the age pattern of mortality in our model is entirely explained by the q(7d) vs. q(7d,5y) contrast. These shifts in the q(x) and ${}_nM_x$ curves in response to changes in k also reflect regularities in our database. They show that a given level of q(5y) can be reached via a variety of routes, depending on a population's unique set of environmental and behavioral conditions. Yet these routes are not unstructured and instead take place within a rather constrained set of possibilities.

As discussed above, almost all data points used for estimating the model fall between k=-1 and +1. This means that predicted values of q(x) using values of k outside that range will represent extrapolations of the model. While the model can certainly tolerate some extrapolation, extrapolating k beyond the range of observed values (a range which spans between -1.1126 and +1.522, as we estimate using a procedure discussed in the next section) should not be performed as they will not have any empirical basis. Moreover, predicted values of q(x) when k<-1.5 will sometimes produce a non-monotonic progression in q(x), which is impossible. As a rule of thumb, users should use the model with k ranging from -1.1 and +1.5.

Estimating the Value of k for a Given Population

Our model can summarize a full set of observed q(x)'s between 0 and 5 years for a given population with only two parameters: q(5y) and k. The first parameter, q(5y), can be directly taken from the observed data. The second parameter k, however, needs to be estimated using model coefficients.

One option consists of finding the value of k which, together with the observed value of q(5y) for a given population i, produces a predicted value of q(x) for a given age $x < 5y$ that exactly matches the observed value of q(x) for that population. This value of k, which we call $k_i(x)$, is given in the following equation, derived from Equations (1) and (2):

$$k_i(x) = \frac{e_i(x)}{\mathrm{v}_x} \qquad (3)$$

where $e_i(x)$ is the difference between the predicted and observed values of q(x) when the prediction is performed with k=0, and v_x is taken from Table 2. Equation (3) implies that a value of k for a given population can be estimated on the basis of only one value of q(x) in addition to q(5y).

The value of k can also be estimated using more than one observed value of q(x) in addition to q(5y). Several approaches are possible in this case. For example, one could simply use the mean or median of the $k_i(x)$ values calculated independently for each age using Equation (3). Another approach consists of finding the value of k which, together with the observed value of q(5y), minimizes the Root Mean Square Error (RMSE) of predicted values of all the q(x) values for that population. To derive the equation for this "best-fitting" value of k, which we denote k_i^*, we take into account the different lengths of the age intervals in the q(x) series by using a weighted least squares solution where the weights w(x) correspond to the length of the previous age interval ending with age x. The solution is given in Equation (4) (see Appendix 1 for more details):

$$k_i^* = \frac{\sum_{x \in X} w(x) \cdot e_i(x) \cdot \mathrm{v}_x}{\sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2} \qquad (4)$$

Compared to the solution based on averages of $k_i(x)$ values, this approach minimizes the uncertainty about the predictions of the model. This is a desirable condition, considering our goal to use this model for indirect estimation and for data validation purposes.

Figure 5 uses data from Finland in 1933 to illustrate how the model can fit an actual observed q(x) series using q(5y) and k*. In Panel A, the circles show the observed values of q(x) at different ages, with a q(5y) value of 107 per 1000. Predicted values of q(x) using the log-quadratic model with this value of q(5y) and k=0 show a certain amount of prediction error. These prediction errors are minimized by calculating the value of k* (=0.99 in this example) using Equation 4. The two entry parameters for Finland in the log-quadratic model are q(5y)=0.107 and k*=0.99, producing a series of predicted q(x)'s that fit the observed data remarkably well, with a RMSE of 1.8%. Figure 5 (Panel B) also shows how the model fits the observed $_nM_x$ series.

--- Figure 5 about here ---

The approach discussed above uses q(5y) as the first entry point, and one or several intermediate q(x) values as additional information for estimating k. For certain applications, it may be desirable to fit the model with input death probabilities that do not start at age 0 and/or do not end at age five years. One example of such configuration is when the only available input values are observed values of q(28d) and q(12m). In some other applications, it may be useful to estimate the model parameters after excluding information at neonatal ages, for example due to concerns about the quality of the data at these ages. In that case, q(28d,5y), rather than q(5y), would be a preferable input parameter. Another situation is when the available input values are mortality rates ($_nM_x$), rather than probabilities, over age groups that do not conform with the model's harmonized age groups. For all these more complex applications, estimating the model parameters cannot be performed using the method described above because of non-linearities in the system of equations. These applications can be resolved using simple iterative procedures, or using a more general approach based on the method of Lagrange. This more general approach is described in Appendix 2.

Our log-quadratic model is a two-dimensional model, but it can be reduced to one dimension assuming k=0. In that case, any single mortality indicator within the 0-5 age range will be associated with one value of q(5y), and a full mortality schedule can be predicted using that q(5y) value and k=0. This corresponds to the model's average prediction in the database given the chosen predictor. In order to take advantage of the two-dimensional feature of the model, at least two input mortality values are necessary. However not all pairs of mortality indicators within the 0 to 5 age range will provide a solution. As discussed above, the shape of the q(x) function, as summarized by the parameter k, is to a large extent driven by the contrast between mortality before vs. after 28 days (or 7 days when q(5y) reaches low levels). This means that, for example, when the pair of input mortality values are both located within the 28d-5y age range, there may not be a solution for q(5y) and k values that produces an exact match for both input values, indicating in effect that the input information is insufficient for determining the shape parameter k. In this case, the two-dimensional model can be reduced to only one dimension assuming k=0, and the model parameter q(5y) can be estimated using either of the two input values, which in such situations will provide similar results. Among classic mortality indicators including q(7d), q(7d,28d), q(28d), q(28d,12m), q(12m), q(12m,5y), and q(5y), pairs that are both on the same side of the 28 days (or 7 days) threshold will most often not provide enough information for estimating the shape parameter k. The same conclusion applies when using 3 or more indicators and solving for the model parameters by minimizing RMSE: these multiple indicators need to combine mortality information before and after the 28 days (or 7 days) threshold to have enough traction for estimating k. When this is not the case, assuming k=0 will be the preferred solution.

How Does the Log-Quadratic Model fit the U5MD?

We evaluated model fitting as the capacity to make q(x) predictions with minimum RMSE for the country-years included in the final U5MD. In order to prevent overfitting, we split our set of country-years into two random samples: one of 60% of the country-years for estimating the coefficients of the model $\{\mathrm{a}_x, \mathrm{b}_x, \mathrm{c}_x, \mathrm{v}_x\}$ and another 40% for evaluating the error of the prediction. We first estimated prediction errors taking q(5y) as the only entry parameter in the model, assuming k=0. We then estimated how model fitting improves when using a second entry point for estimating the shape parameter k, comparing different choices of entry points for that purpose (q(7d), q(28d), q(3m), q(6m) and q(12m)). Finally, we examined model fitting when k is estimated on the basis of all q(x) values, i.e., using k* in Equation (4).

Table 3 shows the RMSEs for both sexes combined selecting 60% of the country-years for estimation and the remaining 40% for evaluation. We report means of the estimates, after preserving the selection 60-40 for a total of 1,000 random samples (without replacement). Global results from 0 to 5 years were calculated as the weighted average of the RMSEs at different ages using the same age weights used in Equation (4). The overall adjustment of the model is satisfactory even if a value of $k = 0$ is assumed, with an RMSE of only 4.04%. Choosing a second entry point and estimating the corresponding value of k improves fit substantially, with the largest improvement occurring with q(3m) as second entry point (RMSE=1.91% for both sexes combined). As expected, best results are obtained when estimating k optimally using k* based on all observed q(x)'s. Interestingly, this optimal solution is not substantially different, in terms of RMSE, from the one using q(3m) for estimating k. Table 3 also shows RMSE when focusing on specific q(x) outcomes: neonatal and infant mortality, i.e., q(28d) and q(12m), respectively. The RMSE's are higher in that case, in part because the global RMSE estimates include values of q(x) at higher ages which have smaller relative prediction errors. Nonetheless the results show that these indicators are relatively well predicted, with predictions that improve overall with the inclusion of the second parameter k.

--- Table 3 about here ---

We also evaluated the performance of the model for predicting mortality outcomes based on q(28d,5y). As mentioned earlier, this indicator excludes mortality information during the neonatal period, making it a useful predictor of neonatal mortality and other under-5 mortality indicators when there are concerns about undercount of deaths at neonatal ages in a given population. Indeed, in such situations, the model's entry point cannot be q(5y), because that indicator is itself affected by undercount of neonatal deaths. Also, estimating k will be problematic, because k is determined to a large extent by the contrast between mortality before vs. after 28 days, which is missing in this configuration.

Predicting a full q(x) schedule in this case can be done assuming k=0. This implies finding the level of q(5y) that matches the observed level of q(28d,5y) when k=0, using either simple iteration or the Lagrange option discussed on Appendix 2. The last row of Table 3 shows the RMSE of q(28d) and other mortality outcomes, here also selecting 60% of the country-years in the database for estimation and the remaining 40% for evaluation. Focusing on q(28d), RMSE are substantially higher than when using q(5y) as a predictor (31.50% vs. 14.48%). This is expected given that q(60m) is to a large extent determined by the level of q(28d), making it easier to predict q(28d) on the basis of q(60m) than on the basis of q(28d,5y). RMSE for other mortality outcomes including q(5y) are substantially lower, due to the overlap in this case between predictor (q(28d,5y) and predicted (q(5y)) indicators. A practical example of using the log-quadratic model for adjusting neonatal mortality based on VR data from Jordan is provided later in the paper. (Note that of all mortality indicators between 28 days and 5 years, q(28d,5y) is the one that produces the smallest predictions errors in q(28d) when assuming k=0. We thus recommend using it when available. Alternatively, neonatal mortality can be predicted using other mortality indicators after 28d, such as q(28d,12m) or q(12m,5y). In that case prediction errors will be slightly higher: 34.41% with q(28d,12m) and 31.79% with q(12m,5y) vs. 31.50% with q(28d,5y).)

Estimating Uncertainty in Predicted q(x) Values

Given q(5y) and k, the log-quadratic model predicts a series of q(x) values. These predictions are not perfectly accurate; the model will predict q(x) values with a certain degree of uncertainty that needs to be quantified.

Our strategy for quantifying uncertainty in predicted values of q(x) values is derived from our approach for estimating ${k_i}^*$, the optimal value of k for a given country i. Building on Equation (4), we obtain in Equation (5) an expression for the variance of ${k_i}^*$ in terms of the prediction error e_i (when $k = 0$), the estimated coefficients for modeling the mortality pattern v, and the optimal value of ${k_i}^*$ (see Appendix 1 for more details):

$$\mathrm{Var}\left[{k_i}^*\right] = \frac{22}{21} \cdot \left[\frac{\sum_{x \in X} w(x) \cdot e_i(x)^2}{\sum_{x \in X} w(x) \cdot \mathrm{v}_x{}^2} - {k_i}^{*2}\right]. \qquad (5)$$

Equation (5) shows that the variance of ${k_i}^*$ is an increasing function of the variance of the prediction error but a decreasing function of the absolute value of ${k_i}^*$. In other words, the certainty in the value ${k_i}^*$will depend on the extent to which the coefficients of the model effectively minimize the RMSE of the prediction. This estimated variance around ${k_i}^*$ can then be used for calculating 95% confidence intervals around each q(x) values predicted by the log-quadratic model. This involves calculating predicted values of q(x) in the log-quadratic model using k $\pm$ 1.96 $\sqrt{\mathrm{Var}\left[{k_i}^*\right]}$.

An illustration of this approach for calculating confidence intervals around predicted q(x) values is provided in Figure 6, using data from Belgium in 1949. We chose this example because of its relatively large remaining prediction errors after estimating k* (RMSE=3.6%), making the calculation of confidence intervals particularly relevant. In Figure 6, each predicted q(x) value is presented with its corresponding 95% confidence interval.

--- Figure 6 about here ---

In one-dimensional uses of the model (i.e., assuming k=0), only one mortality indicator is used as an entry point. Uncertainty in k for a given population in that situation does not stem from variations in $k_i(x)$ across age groups, but instead from the overall lack of information about k. In such cases we propose to build confidence intervals around predicted values by examining patterns of prediction errors in the database when assuming k=0 instead of the best-fitting value k*. We find that across all 1,235 country-year of the final U5MD, the central 95% of the distribution of k* lies between -0.6514 and +0.9362. Confidence intervals around predicted values of q(x) when k=0 can be derived using these bounds for k. An application of this approach is discussed in the next section.

Using the Model for Adjusting Under-5 Mortality in Populations with Incomplete or Deficient Data

Our log-quadratic model for under-5 mortality has many practical applications. It can be used, for example, to: (1) smooth noisy age schedules; (2) correct mortality estimates in the presence of age heaping or transfer; or (3) adjust mortality data for underreporting in specific age ranges.

For the first application, we examine the case of age schedules of mortality estimated using full birth histories collected in the DHS. Mortality information based on this type of information is subject to more sampling error than VR-based information due to small sample sizes. This sampling error is particularly visible when examining age-specific deaths rates (${}_nM_x$) over narrow age intervals. The flexible parametric assumptions of the log-quad model can be used to smooth this information: one simply needs to solve for the model's parameters on the basis of the observed q(x) information, and then use these parameters to obtain predicted values of q(x) from which a smoothed ${}_nM_x$ series can be derived.

An illustration of this application is provided in Figure 7, with data from the 2011-12 DHS in Honduras. The left panel shows observed q(x) values as well as q(x) values predicted using the log-quadratic model given the observed q(5y) value of 28 per 1000 and the best-fitting k* value of 0.06. The model fits the q(x) series extremely well, with a RMSE value of 1.8% and narrow confidence intervals. The right panel shows corresponding observed vs. predicted ${}_nM_x$ values, illustrating the use of the loq-quadratic model for smoothing purposes. The confidence intervals around predicted ${}_nM_x$ values are narrower than suggested by the random error in observed ${}_nM_x$ values. This is explained by the fact that these confidence intervals reflect uncertainty in estimating k, assuming a known value of q(5y), while the observed values of ${}_nM_x$ are affected by sampling error arising from small sample sizes in each narrow age interval.

--- Figure 7 about here ---

The second application deals with age heaping correction. In birth histories collected by DHS surveys, ages at death tend to be reported with a certain amount of heaping, most notably at age 12 months. This raises concerns about the quality of DHS-based IMR estimates, since some infant deaths (i.e., at ages less than 12 months) may be misreported as occurring at 12 months and thus erroneously excluded from IMR calculations (Croft et al. 2018). In order to correct for this issue, we suggest fitting the model to observed q(x) points in a DHS survey after excluding ages most likely to be affected by heaping at 12 months due to their proximity, e.g., 8 to 21 months. The idea is to smooth out age heaping around 12 months while preserving the observed value of q(5y) which is not expected to be affected by such age heaping.

To illustrate this application, we show in Figure 8 data from Bolivia’s 1989 DHS. In this example, the observed q(x) points display a large jump between q(12m) and q(13m), illustrating the extent of age heaping for deaths reported at age 12 months. The observed data suggest an IMR level of 90 per 1000, but this value is questionable given the presence of such age heaping.

Figure 8 also shows predicted values of q(x) with the observed q(5y) value of 140 per 1000 and k estimated on the basis of observed q(x) values excluding the problematic ages around 12 months. The model fits the retained points well (RMSE = 2.8%) and predicts an IMR value of 100 per 1000, i.e., 10 points higher than the observed one. In this example, ages at death in the months preceding 12 months appear to be gradually misreported as occurring at 12 months, generating a substantial downward bias in the observed IMR value.

In the third application, we show how the model can be used to adjust mortality information in situations where mortality may be under-reported at some ages for reasons other than age heaping, for example due to undercount of deaths. In this type of situation, it will not be possible to use the reported value of q(5y) as one of the model's entry points, because that value will itself be biased by such under-reporting. However, as explained earlier, the model's parameters can be estimated using entry points over age ranges that may not start at zero and/or may not end at 5. This allows users to estimate the model's parameters on the basis of indicators within the 0-5 age range that may be less affected by under-reporting issues.

We illustrate this type of application using recent (2015) vital registration data from Jordan, a country where VR-based under-5 mortality information appears largely underestimated (United Nations Inter-agency Group for Child Mortality Estimation (UN IGME) 2019b). As is often the case, concerns about undercount are particularly acute for neonatal mortality, as indicated in the Jordan VR data by an unusually low level of neonatal mortality given the observed level of under-5 mortality. We propose here to use the log-quad model for adjusting under-5 mortality in the country using the observed value of q(28d,5y) as the model's entry point. As discussed earlier, this choice is based on the fact that q(28d,5y) is an indicator that remains unbiased in the presence of underreporting of neonatal deaths. Unlike the previous applications, it will not be possible to solve for the model's second dimension k, because as we saw earlier the estimation of k requires entry points situated on both sides of the 28 days threshold. However, assuming k=0, it is possible to solve for the value of q(5y) that corresponds to the observed value q(28d,5y) and then obtain a full series of predicted q(x) values, including neonatal, infant and under-5 mortality rates. We calculated confidence intervals around predicted values using bounds of k varying between -0.6514 and +0.9362 as discussed in the previous section.

Results of this approach, shown in Figure 9, indicate that the model adjusts the VR estimates of neonatal mortality upwards by a factor of more than 2, from 4 to 10 per 1000, producing adjusted levels that are consistent with DHS estimates for the same period. Figure 9 also shows how this adjustment of neonatal mortality affects levels of infant and under-5 mortality. The adjusted level of q(5y) produced by the log-quad model is 17.5 per 1000, more than 50% higher than the unadjusted level and on a par with the DHS estimate. The consistency between our VR-adjusted estimates and the DHS estimates is reassuring about the ability of our approach to correct for deficiencies in the VR data. Confidence intervals between the two approaches have comparable sizes, although they arise from different reasons. In the case of DHS-data, uncertainty reflects sampling error, while in the case of the VR correction the confidence interval reflects the model's prediction error in the neonatal mortality rate when k=0.

-- Figure 9 about here ---

In order to further understand the mechanics of this adjustment, we show in Figure 10 observed vs. predicted values of ${}_nM_x$ in Jordan with a focus on the first 12 months. This figure shows that, while there is close agreement between observed and predicted rates from the second week of life onwards, the model predicts much higher mortality for the first week. This suggests that under-reporting in neonatal mortality in the VR data for Jordan comes primarily from under-reporting during the first week, which is indeed the age range most sensitive to data errors. Overall, this approach offers a promising solution for adjusting VR-based estimates of under-5 mortality in situations where issues of undercount are concentrated at neonatal ages. This solution is particularly useful given the renewed emphasis on using local vital registration information rather than international survey programs as a data source for estimating mortality.

-- Figure 10 about here ---

Discussion

Mortality between 0 and 5 has features that make this age range unique over the human life course, including a particularly fast decline by age during the first weeks and months of life that has been interpreted using evolutionary and selection models (Chu et al. 2008; Lee 2003; Schöley 2019). Over the history of mortality change, populations have experienced large changes in both the level and shape of under-5 mortality in response to epidemiological changes such as a shift from exogenous causes of under-5 death (e.g., infectious and parasitic diseases) to endogenous causes (e.g., congenital malformations, birth injuries) (Drevenstedt et al. 2008; Galley & Woods 1999; Liu et al. 2012; Rao et al. 2011). As a tool for describing and summarizing these regularities, the new model developed in this paper has a number of strengths. First, with 22 age groups between age 0 and 5, our model offers far more detailed age granularity than existing model life tables. This age detail is particularly relevant given the fast changes in age-specific mortality in that age range. Second, the model fits high-quality VR data remarkably well. Thus, the two-dimensional log-quadratic approach is well suited to describe changes in both the level and shape of under-5 mortality observed in the populations represented in our database. Third, our model provides a more flexible choice of predictors, beyond the typical infant vs. child mortality contrast embedded in classic model life tables. This allows users to predict mortality curves using various combinations of predictors depending on data availability and quality. Fourth, our model can be used for various data smoothing and adjustment applications, as shown in our empirical applications. Our application of the model to data from Jordan, in particular, shows how the model can be used for correcting incomplete VR data in situations where under-reporting is concentrated during the neonatal period. Fifth, unlike most model life table approaches, our model provides solutions for estimating confidence intervals around predicted values. Finally, our model is simple and easy to use. The coefficients provided in Table 2 contain all the necessary information for using the model, and most applications can be solved using simple formulas such as Equations (3) and (4).

The model's main limitation is that its empirical basis does not include mortality data from low- and middle-income countries. This means that applications of the model to a low- and middle-income populations need to rely on the assumption that the mortality regularities described by our model, representing mostly the experience of historical and contemporary Western countries, apply to that particular population. Our examples from Honduras, Bolivia and Jordan for recent periods suggest that the model's applicability is broader than the geographical scope of the

U5MD. Indeed, in all three cases, there was a close fit between observed and predicted values of q(x) for the ages used as a basis for the prediction.

There are cases, however, where the model is clearly not able to reproduce the observed age patterns for reasons that appear unrelated to data quality issues in the observed data. The most extreme cases are populations that exhibit a large age-specific reversal in mortality around age 6 months, as was observed for example in the Niakhar surveillance site in Senegal in the 1960s and 1970s (Abdullah et al. 2007; Cantrelle & Leridon 1971; Delaunay et al. 2001; Lalou & LeGrand 1996). This unusual age pattern, which has been attributed to a combination of factors including inadequate weaning foods (Cantrelle & Leridon 1971; Garenne 1982), is absent from the Western experience, and thus our log-quadratic model is not able to reproduce it. Outside these extreme cases, many sub-Saharan African populations tend to display an unusually late age pattern of under-5 mortality (Guillot et al. 2012), which is not well fitted by the log-quadratic model (Romero et al. 2019). As an illustration, we show in Figure 11 observed q(x) values from the 2011-12 DHS for Senegal against log-quadratic predictions given the same level of q(5y), with k varying between -1 and +1. Clearly the log-quadratic model is not able to reproduce this age pattern, which combines an unusually high level of neonatal mortality (associated with an "early" pattern of under-5 mortality in the log-quadratic model) with unusually low values of q(x) at later ages (associated with a "late" pattern). This lack of fit shows that while the log-quadratic model can be applied to various non-Western populations, it cannot be used indiscriminately everywhere.

In a recent paper, Mejía-Guevara et al. (2019) specifically modeled age patterns of under-5 mortality in sub-Saharan Africa using DHS data calibrated on estimates from the United Nations

Inter-agency Group for Child Mortality Estimation (UN IGME) (2019b). This study, like ours, recognizes the importance of age patterns of mortality as a device for mortality estimation, but it pursues objectives that are substantially different from ours, and thus is not directly comparable. Its goal is primarily to smooth and forecast existing data on under-5 mortality by detailed age; by contrast, our study follows a model life table approach, which consists of extracting regularities from a reference dataset via coefficients that may then be used for evaluating and correcting data in populations not included in that dataset. Nonetheless, Mejia-Guevara et al.'s (2019) study raises the question of whether DHS data may be used as a source for modeling age patterns in low- and middle-income countries, including sub-Saharan Africa. In our study, we chose not to include DHS data due to data quality concerns that are particularly consequential given the specific goals of our model, including age heaping and concerns about the quality of the reporting of neonatal deaths (Helleringer et al. 2020). This does not mean that our model's inability to fit patterns such as the one shown in Figure 11 for Senegal is indicative of data errors in the DHS. There are many reasons to believe that age patterns of under-5 mortality in many sub-Saharan African populations are truly different from those observed in Western countries. However, we believe that our goal to derive a model that can be used as a reference for data evaluation and correction requires a thorough evaluation of all the available sources of under-5 mortality information in low- and middle-income countries, an exercise that is beyond the scope of this paper. We provide here a model that describes age patterns based on gold-standard, newly compiled vital registration data spanning a large number of countries and time periods. Nonetheless more research is needed augment the geographical scope and generalizability of this model.

Conclusion

This paper proposes a new model for summarizing regularities about how under-5 mortality is distributed by detailed age. This model is based on a newly compiled database that contains under-5 mortality information by detailed age in countries with high-quality vital registration systems, covering a wide array of mortality levels and patterns. This model uses a log-quadratic approach, predicting a full mortality schedule between age 0 and 5 on the basis of only 1 or 2 parameters.

Results show that our model is able to accurately describe variations in both the level and shape of under-5 mortality across a variety of contexts. We believe that our model, with its innovative features relative to existing models, contributes to better estimating and understanding levels and age patterns of under-5 mortality. Future research should focus on increasing the geographical scope of the model by gathering the best possible data on under-5 mortality by detailed age in low- and middle-income countries.

References

Abdullah, S., Adazu, K., Masanja, H., Diallo, D., Hodgson, A., Ilboudo-Sanogo, E., et al. (2007). Patterns of age-specific mortality in children in endemic areas of Sub-Saharan Africa. *The American Journal of Tropical Medicine and Hygiene, 77*(6 Suppl), 99-105.

Agorastakis, M., Jdanov, D., & Grigoriev, P. (2017). *About mortality data for Greece*. Retrieved from https://www.mortality.org/

Barbieri, M., Wilmoth, J. R., Shkolnikov, V. M., Glei, D., Jasilionis, D., Jdanov, D., et al. (2015). Data resource profile: The Human Mortality Database (HMD). *International Journal of Epidemiology, 44*(5), 1549-1556. doi:10.1093/ije/dyv105

Bourgeois-Pichat, J. (1951). La mesure de la mortalité infantile. I. Principes et méthodes. *Population (French Edition), 6*(2), 233-248. doi:10.2307/1524151

Cantrelle, P., & Leridon, H. (1971). Breast feeding, mortality in childhood and fertility in a rural zone of Senegal. *Population Studies, 25*(3), 505-533. doi:10.1080/00324728.1971.10405821

Chu, C. Y. C., Chien, H.-K., & Lee, R. D. (2008). Explaining the optimality of u-shaped age-specific mortality. *Theoretical Population Biology, 73*(2), 171-180. doi:10.1016/j.tpb.2007.11.005

Clark, S. J. (2019). A general age-specific mortality model with an example indexed by child mortality or both child and adult mortality. *Demography, 56*(3), 1131-1159. doi:10.1007/s13524-019-00785-3

Coale, A. J., & Demeny, P. G. (1966). *Regional model life tables and stable populations*. Princeton, N.J.,: Princeton University Press.

Coale, A. J., Demeny, P. G., & Vaughan, B. (1983). *Regional model life tables and stable populations* (2nd ed.). New York: Academic Press.

Croft, T. N., Marshall, A. M. J., Allen, C. K., & et al. (2018). *Guide to DHS statistics*. Retrieved from https://dhsprogram.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_DHS-7.pdf

Delaunay, V., Etard, J.-F., Préziosi, M.-P., Marra, A., & Simondon, F. (2001). Decline of infant and child mortality rates in rural Senegal over a 37-year period (1963-1999). *International Journal of Epidemiology, 30*(6), 1286-1295. doi:10.1093/ije/30.6.1286

Drevenstedt, G. L., Crimmins, E. M., Vasunilashorn, S., & Finch, C. E. (2008). The rise and fall of excess male infant mortality. *Proceedings of the National Academy of Sciences of the United States of America, 105*(13), 5016-5021. doi:10.1073/pnas.0800221105

Galley, C., & Woods, R. (1998). Reflections on the distribution of deaths in the first year of life. *Population, 53*(5), 921-946. doi:10.2307/1534830

Galley, C., & Woods, R. (1999). On the distribution of deaths during the first year of life. *Population: An English Selection, 11*(1), 35-59.

Garenne, M. L. (1982). *Variations in the Age Pattern of Infant and Child Mortality with Special Reference to a Case Study in Ngayokheme (Rural Senegal).* (Ph.D. 8307314), University of Pennsylvania, United States -- Pennsylvania. ProQuest Dissertations & Theses (PQDT) database.

Gourbin, G., & Masuy-Stroobant, G. (1995). Registration of vital data: Are live births and stillbirths comparable all over Europe? *Bulletin of the World Health Organization, 73*(4), 449-460.

Guillot, M., Gerland, P., Pelletier, F., & Saabneh, A. (2012). Child mortality estimation: A global overview of infant and child mortality age patterns in light of new empirical data. *PLOS Medicine, 9*(8), e1001299. doi:10.1371/journal.pmed.1001299

Helleringer, S., Liu, L., Chu, Y., Rodrigues, A., & Fisker, A. B. (2020). Biases in Survey Estimates of Neonatal Mortality: Results from a Validation Study in Urban Areas of Guinea-Bissau. *SocArXiv. March 20*. doi:10.31235/osf.io/qx2kn

Hill, K. (1995). Age patterns of child mortality in the developing world. *Population Bulletin of the United Nations*(39), 112-132.

Hug, L., Alexander, M., You, D., & Alkema, L. (2019). National, regional, and global levels and trends in neonatal mortality between 1990 and 2017, with scenario-based projections to 2030: a systematic analysis. *The Lancet Global Health, 7*(6), e710-e720. doi:https://doi.org/10.1016/S2214-109X(19)30163-9

Knodel, J., & Kintner, H. (1977). Impact of Breast Feeding Patterns on Biometric Analysis of Infant-Mortality. *Demography, 14*(4), 391-409. doi:Doi 10.2307/2060586

Lalou, R., & LeGrand, T. K. (1996). Child mortality in towns and villages in the Sahel region. *Population, 51*(2), 329-351. doi:10.2307/1534584

Lantoine, C., & Pressat, R. (1984). New aspects of infant-mortality. *Population, 39*(2), 253-264. doi:10.2307/1532294

Lawn, J. E., Osrin, D., Adler, A., & Cousens, S. (2008). Four million neonatal deaths: Counting and attribution of cause of death. *Paediatric and Perinatal Epidemiology, 22*(5), 410-416. doi:10.1111/j.1365-3016.2008.00960.x

Lee, R. D. (2003). Rethinking the evolutionary theory of aging: Transfers, not births, shape senescence in social species. *Proceedings of the National Academy of Sciences of the United States of America, 100*(16), 9637-9642. doi:10.1073/pnas.1530303100

Lee, R. D., & Carter, L. R. (1992). Modeling and forecasting U.S. Mortality *Journal of the American Statistical Association, 87*(419), 659-671.

Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., et al. (2012). Global, regional, and national causes of child mortality: An updated systematic analysis for 2010 with time trends since 2000. *The Lancet, 379*(9832), 2151-2161. doi:10.1016/S0140-6736(12)60560-1

Lynch, K. A., Greenhouse, J. B., & Brändström, A. (1998). Biometric modeling in the study of infant mortality: Evidence from nineteenth-century Sweden. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 31*(2), 53-64. doi:10.1080/01615449809601188

Manfredini, M. (2004). The Bourgeois-Pichat's biometric method and the influence of climate: New evidences from late 19th-century Italy. *Social Biology, 51*(1-2), 24-36. doi:10.1080/19485565.2004.9989081

Mejía-Guevara, I., Zuo, W., Bendavid, E., Li, N., & Tuljapurkar, S. (2019). Age distribution, trends, and forecasts of under-5 mortality in 31 Sub-Saharan African countries: A modeling study. *PLOS Medicine, 16*(3), e1002757. doi:10.1371/journal.pmed.1002757

Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., de Savigny, D., Lozano, R., et al. (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. *The Lancet, 386*(10001), 1395-1406. doi:10.1016/S0140-6736(15)60171-4

Murray, C. J. L., Ferguson, B. D., Lopez, A. D., Guillot, M., Salomon, J. A., & Ahmad, O. (2003). Modified logit life table system: Principles, empirical validation, and application. *Population Studies, 57*(2), 165-182. doi:10.1080/0032472032000097083

Rao, C., Adair, T., & Kinfu, Y. (2011). Using historical vital statistics to predict the distribution of under-five mortality by cause. *Clinical Medicine & Research, 9*(2), 66-74. doi:10.3121/cmr.2010.959

Romero, J., Verhulst, A., & Guillot, M. (2019). *Estimating IMR from DHS full birth histories in the presence of age heaping.* Paper presented at the Meeting of the Population Association of America, Austin, TX.

Schöley, J. (2019). *The Age-Trajectory of Infant Mortality in the United States: Parametric Models and Generative Mechanisms*. Paper presented at the 2019 Meetings of the Population Association of America (PAA), Austin, TX.

Steffen, M. (1990). A simple method for monotonic interpolation in one dimension. *Astronomy and Astrophysics, 239*, 443-450.

United Nations. (1988). *MortPak-lite -- the United Nations software package for mortality measurement : Interactive software for the IBM-PC and compatibles*. New York: United Nations.

United Nations. (2011). *The millennium development goals report 2011*. New York: United Nations.

United Nations Department of Social Affairs - Population Division. (1954). *Foetal, infant and early childhood mortality*. New York: United Nations.

United Nations Inter-agency Group for Child Mortality Estimation (UN IGME). (2019a). *Explanatory notes: Child mortality trend series to 2018*. Retrieved from https://childmortality.org/wp-content/uploads/2019/09/UNIGME-Explanatory-Notes_ENGLISH.pdf

United Nations Inter-agency Group for Child Mortality Estimation (UN IGME). (2019b). *Levels & trends in child mortality: Report 2019, estimates developed by the UN inter-agency group for child mortality estimation*. Retrieved from https://www.unicef.org/media/60561/file/UN-IGME-child-mortality-report-2019.pdf

United Nations Population Division. (1982). *Model life tables for developing countries*. New York: United Nations.

United Nations Statistical Office. (1955). *Handbook of vital statistics methods*. New York: Statistical Office of the United Nations. Department of Economic and Social Affairs.

Wang, H., Bhutta, Z. A., Coates, M. M., Coggeshall, M., Dandona, L., Diallo, K., et al. (2016). Global, regional, national, and selected subnational levels of stillbirths, neonatal, infant, and under-5 mortality, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet, 388*(10053), 1725-1774. doi:https://doi.org/10.1016/S0140-6736(16)31575-6

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., & Sawyer, C. (2012). A flexible two-dimensional mortality model for use in indirect estimation. *Population Studies, 66*(1), 1-28. doi:10.1080/00324728.2011.611411

Wilmoth, J. R. (1990). Variation in vital rates by age, period, and cohort. *Sociological Methodology, 20*, 295-335.

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D. A., & Riffe, T. (2017). *Methods protocol for the human mortality database*. Retrieved from https://www.mortality.org/Public/Docs/MethodsProtocol.pdf

You, D., Hug, L., Ejdemyr, S., Idele, P., Hogan, D., Mathers, C., et al. (2015). Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: A systematic analysis by the UN inter-agency group for child mortality estimation. *The Lancet, 386*(10010), 2275-2286. doi:10.1016/s0140-6736(15)00120-8

Table 1: List of country-years in the original Under-Five Mortality Database (U5MD) and in the final U5MD used for modeling purposes

	Original U5MD		*Excluded country-years due to insufficient age breakdowns*		*Excluded country-years due to reversals in the $_7M_{0(d)}$ vs. q(28d,5y) relationship*		*Final U5MD for modeling purposes*	
Country	**Year interval**	**n**	**Year interval**	**n**	**Year interval**	**n**	**Year interval**	**n**
Australia	1921-2014	93					1921-2014	93
Austria	1970-2016	46					1970-2016	46
Belgium	1841-2014	98	1841-1861	21	1878-1945	29	1946-2014	48
Canada	1929-2006	71					1929-2006	71
Chile	1992-2007	14					1992-2007	14
Denmark	1890-2015	120	1890-1920	30	1921-1928	8	1929-2015	82
Finland	1878-2015	124			1878-1921	44	1926-2015	80
France	1855-2015	138			1855-1952	87	1953-2015	51
Germany	1991-2015	19					1991-2015	19
West Germany	1956-1990	24					1956-1990	24
Ireland	1970-2011	39					1970-2011	39
Israel	1983-2016	33					1983-2016	33
Italy	1872-2013	99	1872-1889	18	1926-1945	15	1946-2013	66
Japan	1947-2014	53					1947-2014	53
Netherlands	1850-2008	49	1850-1864	15			1970-2008	34
New Zealand	1970-2013	43	1972	1			1970-2013	42
Norway	1876-2012	127			1876-1935	53	1936-2012	74
Portugal	1940-2015	61	1940-1954	14	1955-1970	7	1971-2015	40
South Korea	2004-2015	12					2004-2015	12
Spain	1976-2013	30					1976-2013	30
Sweden	1891-2012	121			1891-1933	43	1934-2012	78
Switzerland	1877-2016	60					1877-2016	60
UK	1982-2012	25					1982-2012	25
England and Wales	1905-1985	81			1905-1936	32	1937-1985	49
United States	1933-2015	72					1933-2015	72
Total	**1841-2016**	**1,652**	**1841-1972**	**99**	**1855-1974**	**318**	**1877-2016**	**1,235**

Table 2: Coefficients of the log-quadratic model estimated with the final U5MD, by sex and for both sexes combined

	Female				*Male*				*Both sexes*			
x	a_x	b_x	c_x	v_x	a_x	b_x	c_x	v_x	a_x	b_x	c_x	v_x
7d	-3.5756	-0.2551	-0.1406	-0.4749	-3.1419	-0.1484	-0.1378	-0.4708	-3.3460	-0.2033	-0.1399	-0.4745
14d	-2.9909	-0.0176	-0.1115	-0.4226	-2.6246	0.0735	-0.1087	-0.4228	-2.7955	0.0276	-0.1106	-0.4237
21d	-2.5866	0.1510	-0.0915	-0.3909	-2.2642	0.2317	-0.0889	-0.3930	-2.4135	0.1916	-0.0905	-0.3927
28d	-2.3579	0.2425	-0.0807	-0.3672	-2.0521	0.3207	-0.0779	-0.3703	-2.1931	0.2822	-0.0796	-0.3695
2m	-1.8380	0.4305	-0.0594	-0.2865	-1.5462	0.5153	-0.0543	-0.2905	-1.6780	0.4751	-0.0569	-0.2884
3m	-1.6142	0.4897	-0.0536	-0.2324	-1.3426	0.5698	-0.0486	-0.2345	-1.4654	0.5318	-0.0511	-0.2329
4m	-1.4530	0.5294	-0.0500	-0.1946	-1.2041	0.6028	-0.0454	-0.1953	-1.3171	0.5677	-0.0478	-0.1940
5m	-1.3079	0.5718	-0.0457	-0.1685	-1.0778	0.6406	-0.0412	-0.1665	-1.1818	0.6080	-0.0435	-0.1662
6m	-1.1751	0.6131	-0.0416	-0.1482	-0.9586	0.6802	-0.0369	-0.1464	-1.0556	0.6489	-0.0392	-0.1461
7m	-1.0544	0.6520	-0.0377	-0.1330	-0.8512	0.7172	-0.0329	-0.1305	-0.9421	0.6869	-0.0353	-0.1304
8m	-0.9529	0.6850	-0.0344	-0.1205	-0.7587	0.7500	-0.0293	-0.1178	-0.8451	0.7200	-0.0318	-0.1178
9m	-0.8634	0.7146	-0.0315	-0.1092	-0.6763	0.7798	-0.0261	-0.1076	-0.7587	0.7502	-0.0286	-0.1073
10m	-0.7803	0.7433	-0.0285	-0.1007	-0.6063	0.8049	-0.0234	-0.0987	-0.6831	0.7769	-0.0258	-0.0986
11m	-0.7077	0.7688	-0.0258	-0.0931	-0.5453	0.8273	-0.0209	-0.0916	-0.6167	0.8009	-0.0232	-0.0911
12m	-0.6436	0.7914	-0.0235	-0.0868	-0.4891	0.8484	-0.0186	-0.0857	-0.5570	0.8226	-0.0209	-0.0850
15m	-0.4887	0.8459	-0.0178	-0.0710	-0.3554	0.8985	-0.0129	-0.0706	-0.4130	0.8753	-0.0151	-0.0696
18m	-0.3857	0.8831	-0.0136	-0.0598	-0.2724	0.9293	-0.0093	-0.0593	-0.3213	0.9089	-0.0112	-0.0583
21m	-0.3107	0.9104	-0.0104	-0.0510	-0.2160	0.9499	-0.0067	-0.0502	-0.2569	0.9324	-0.0084	-0.0495
24m	-0.2507	0.9311	-0.0081	-0.0429	-0.1698	0.9657	-0.0047	-0.0424	-0.2044	0.9506	-0.0062	-0.0417
36m	-0.1254	0.9674	-0.0040	-0.0204	-0.0759	0.9907	-0.0015	-0.0221	-0.0968	0.9806	-0.0026	-0.0211
48m	-0.0466	0.9897	-0.0013	-0.0083	-0.0302	0.9974	-0.0005	-0.0086	-0.0370	0.9941	-0.0008	-0.0084
60m	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000

Table 3: Root Mean Square Error (RMSE) of predicted q(x)'s using the log-quadratic model applied to the final U5MD with various combinations of outcomes and entry points for estimating k, both sexes combined.

		RMSE for the following outcomes:			
Entry point(s)		*all q(x)*	*q*(*28d*)	*q*(*12m*)	*q(5y)*
q(*5y*) only, k=0		0.0404	0.1448	0.0463	0.0000
q(*5y*) and	*q*(*7d*)	0.0248	0.0452	0.0387	0.0000
	q(*28d*)	0.0226	0.0000	0.0376	0.0000
	q(*3m*)	0.0191	0.0549	0.0300	0.0000
	q(*6m*)	0.0226	0.1103	0.0183	0.0000
	q(*12m*)	0.0325	0.1634	0.0000	0.0000
	all *q*(*x*) *	0.0176	0.0571	0.0253	0.0000
q(28d, 5y) only, k=0		0.1908	0.3150	0.2022	0.1687

Reported values correspond to the mean of 1,000 random samples: 60% of the life tables were used for estimation and 40% for evaluation.

RMSE calculated from the relative error of 494 life tables (40% of sample).

* Using k = k* (Equation (4))

Figure 1: Relationship between age-specific mortality rates (${}_{n}M_{x}$) and the probability of dying between age 28 days and 5 years (q(28d,5y)) for each of the first four weeks of life (${}_{7}M_{0(d)}$, ${}_{7}M_{7(d)}$, ${}_{7}M_{14(d)}$ and ${}_{7}M_{21(d)}$) in the original Under-5 Mortality Database (U5MD), both sexes combined.

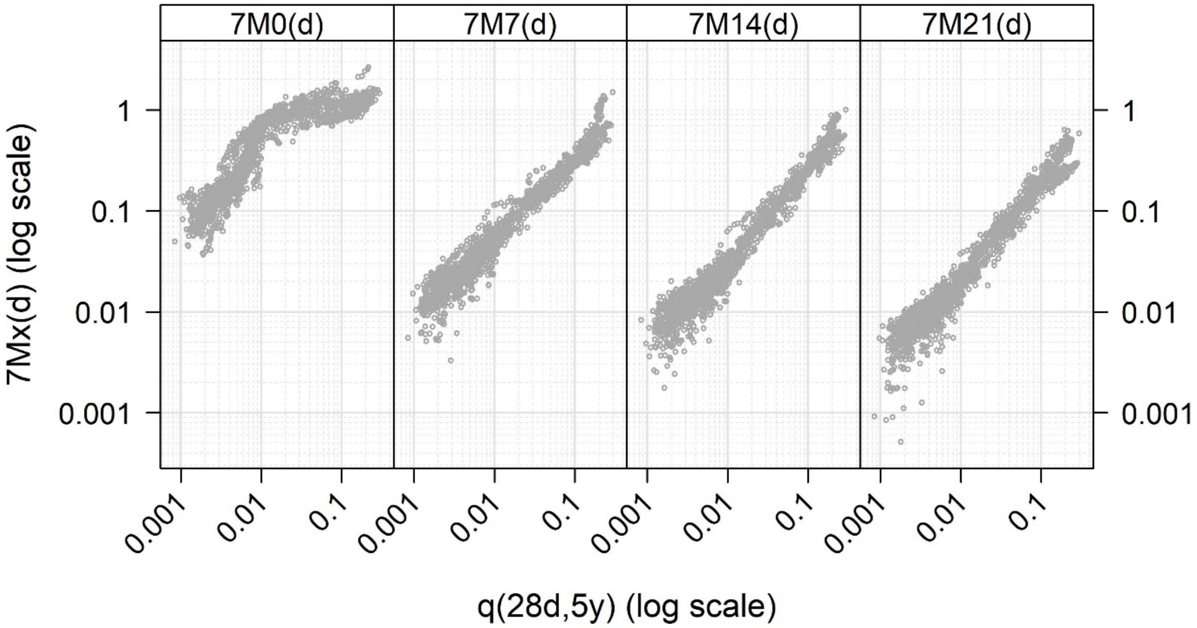


Figure 2: Relationship between q(x) and q(5y) for x=7d, 28d and 12m, with observed values in the final U5MD vs. values predicted using the log-quadratic model with k=0±1

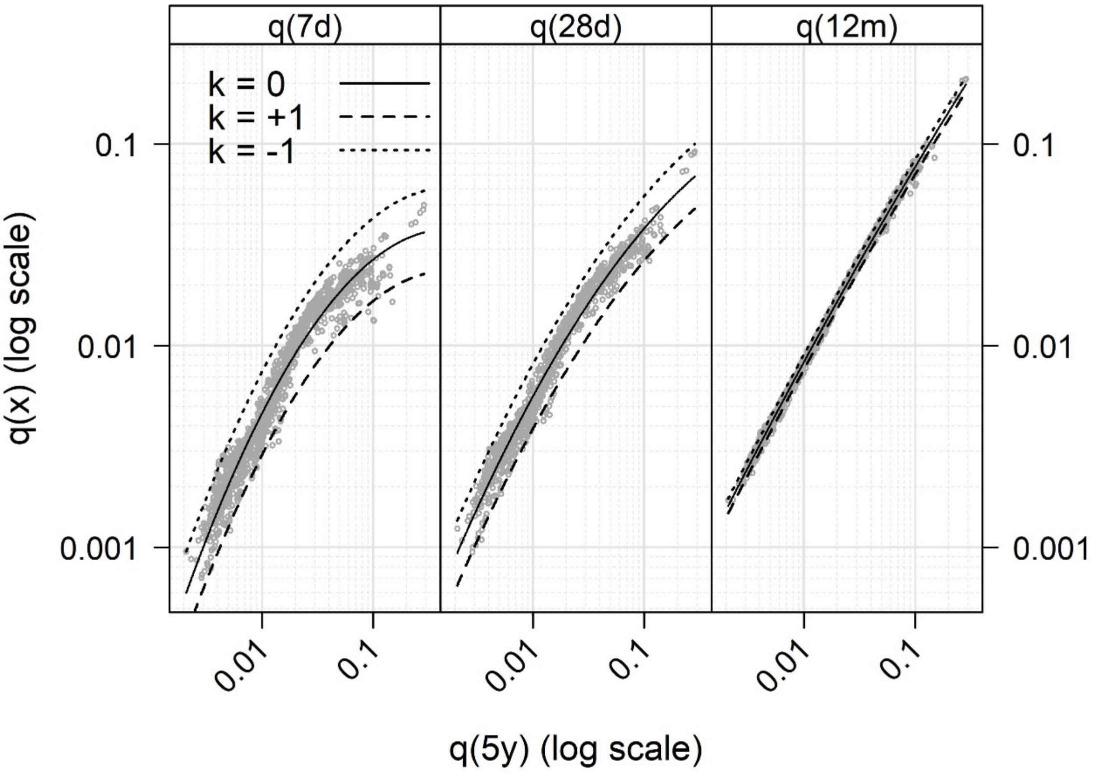
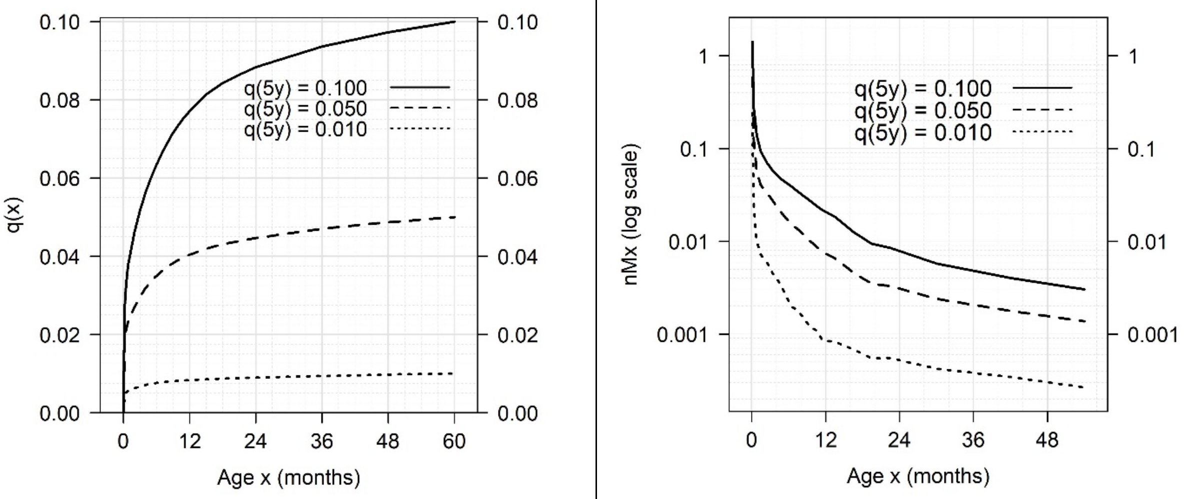


Figure 3: Effect of varying q(5y) on q(x) and ${}_{n}M_{x}$ when k=0 in the log-quadratic model

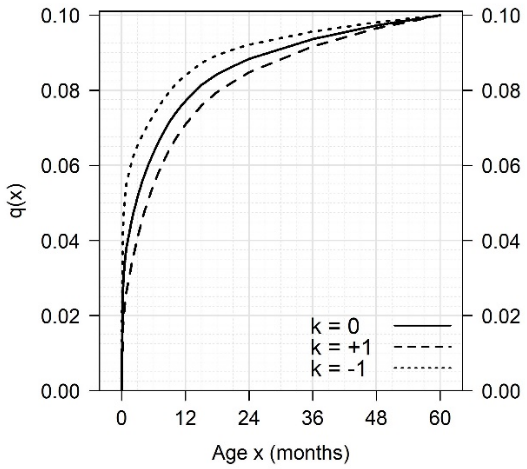


Panel A: Effect of varying q(5y) on q(x)

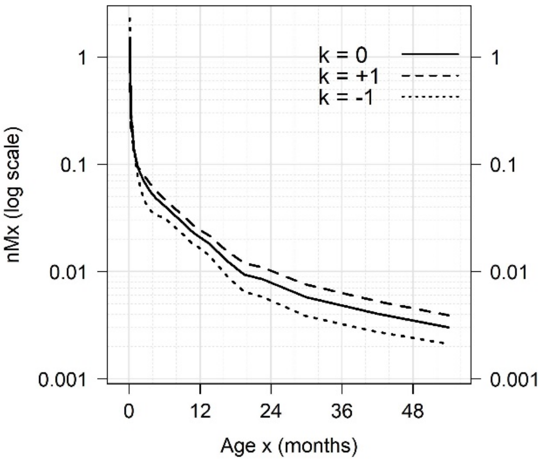
Panel B: Effect of varying q(5y) on ${}_{n}M_{x}$

Figure 4: Effect of varying k on q(x) and $_{n}M_{x}$ when q(5y)=100 p.1000 in the log-quadratic model

Panel A: Effect of varying k on q(x)



Panel B: Effect of varying k on $_{n}M_{x}$ (between 0 to 5 years)



Panel C: Effect of varying k on $_{n}M_{x}$ (between 0 and 3 months)

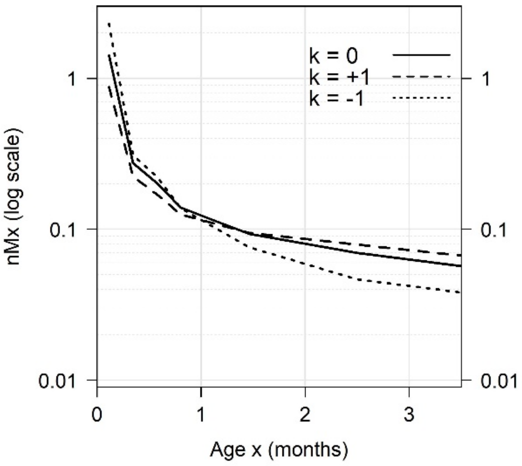
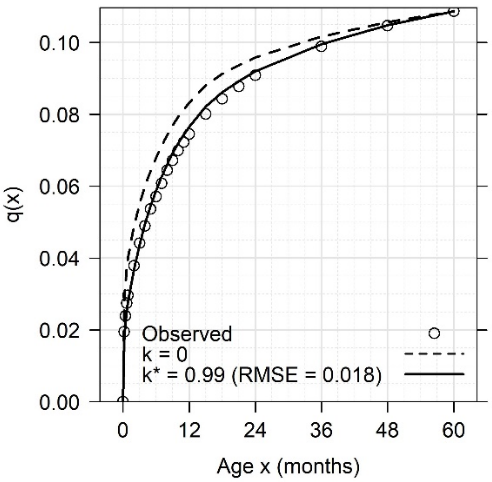


Figure 5: Observed and predicted values of q(x) and $_{n}M_{x}$ for Finland, 1933, both sexes.

Panel A: q(x)



Panel B: $_{n}M_{x}$

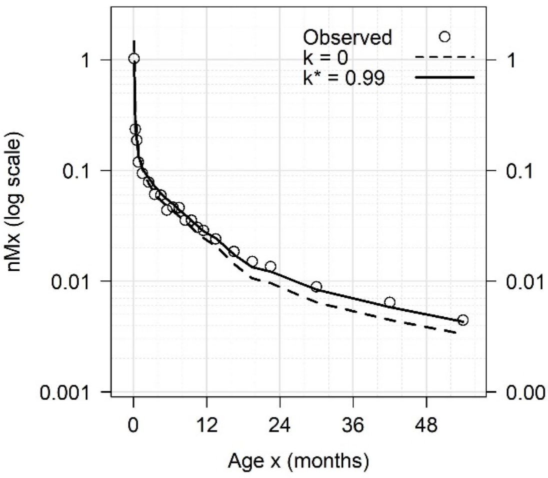
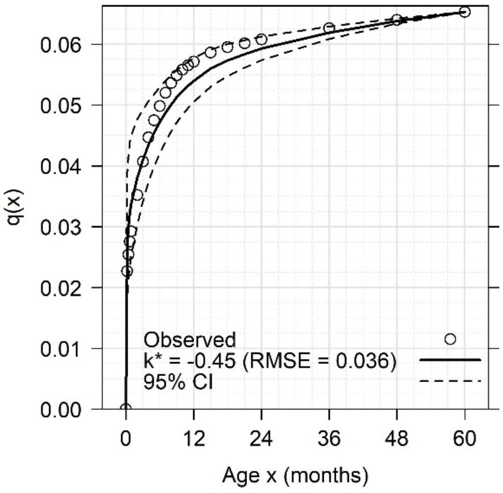


Figure 6: Observed and predicted values of q(x) and ${}_nM_x$ for Belgium, 1949, both sexes, with 95% confidence intervals

Panel A: q(x)



Panel B: ${}_nM_x$

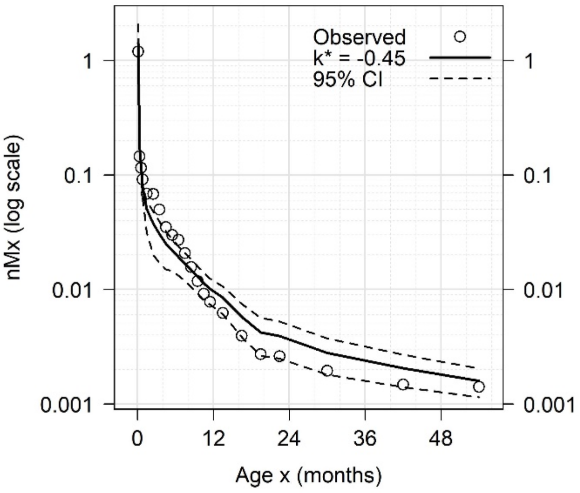
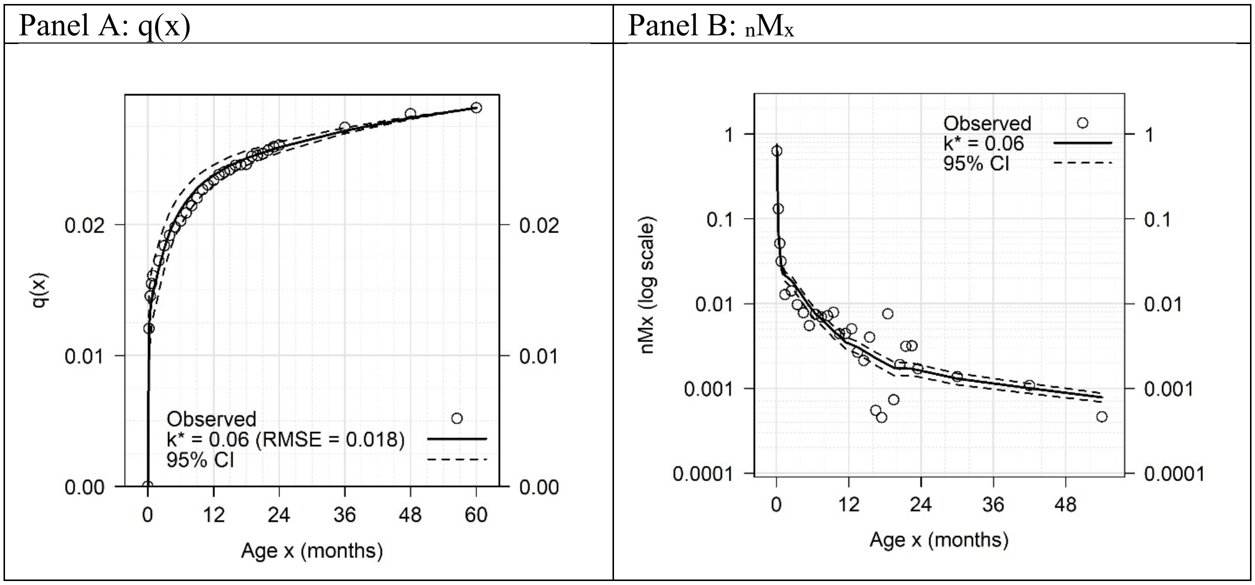


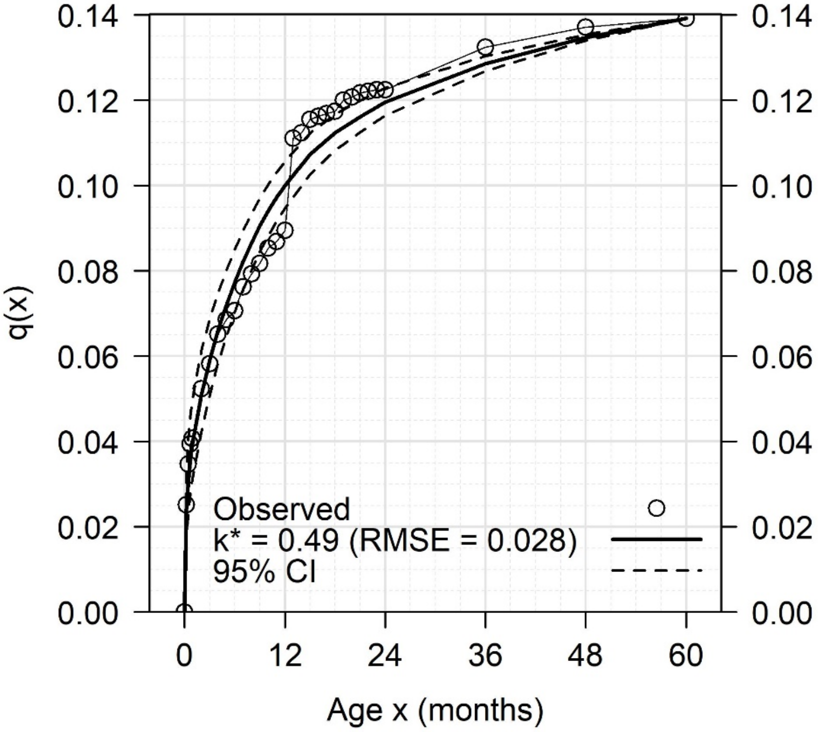
Figure 7: Observed and predicted values of q(x) and ${}_nM_x$ for Honduras 2011-12 DHS survey, both sexes.



Panel A: q(x)

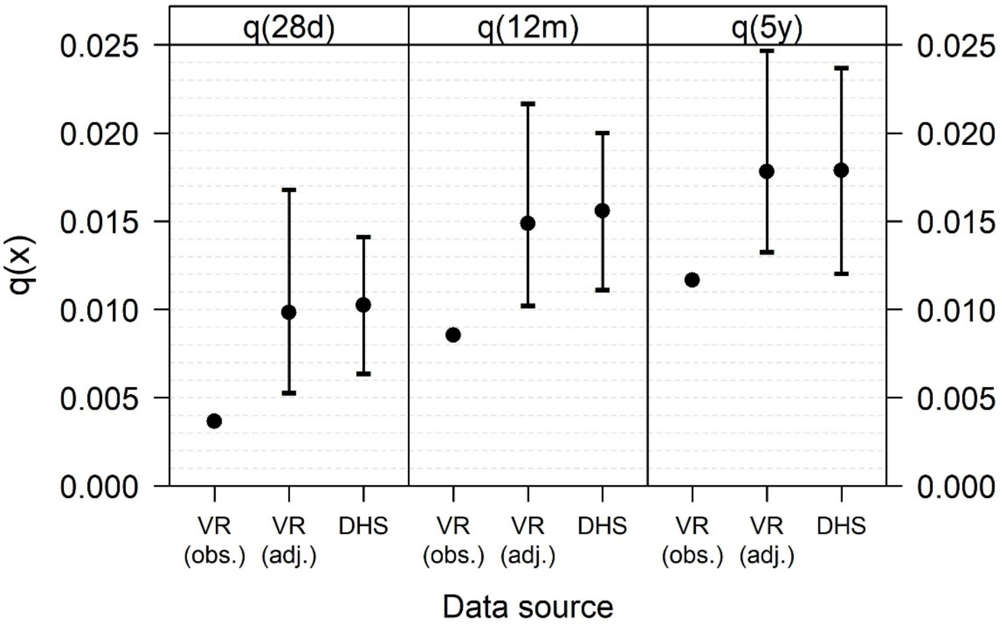
Panel B: ${}_nM_x$

Figure 8: Observed and predicted values of q(x) in the 1989 DHS for Bolivia, both sexes.



Note: k* is estimated here on the basis of observed q(x) points excluding ages between 8 months and 21 months.

Figure 9: Levels of neonatal (q(28d)), infant (q(12m)) and under-five (q(5y)) mortality rates for Jordan in 2015, both sexes, with 95% confidence intervals.

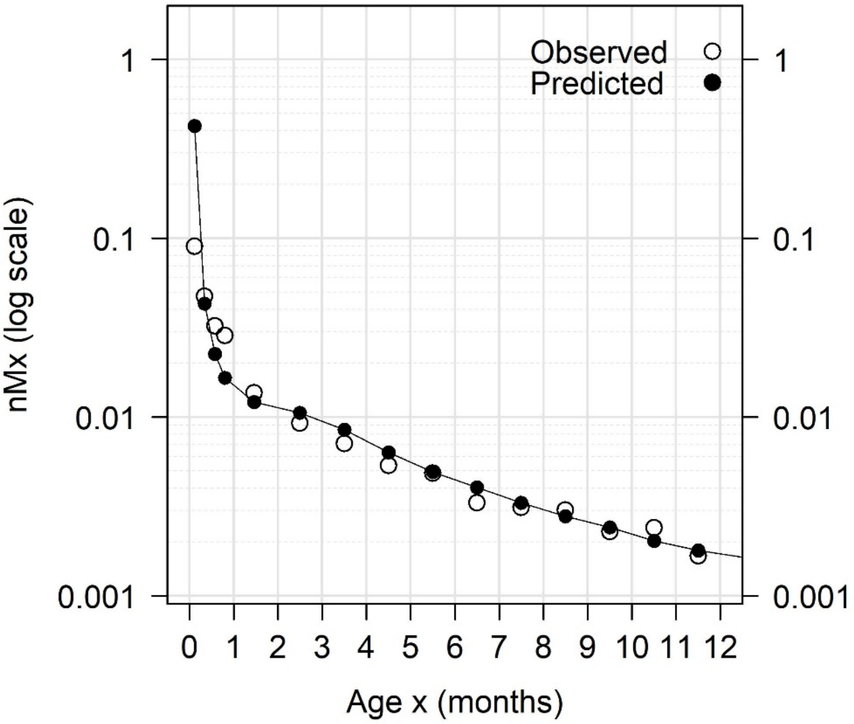


Legend: VR=Vital registration (VR); DHS=Demographic and Health Survey;

obs. = unadjusted VR values

adj. = adjusted VR values using the log-quadratic model with the observed VR-based value of q(28d,5y) and k=0 as inputs

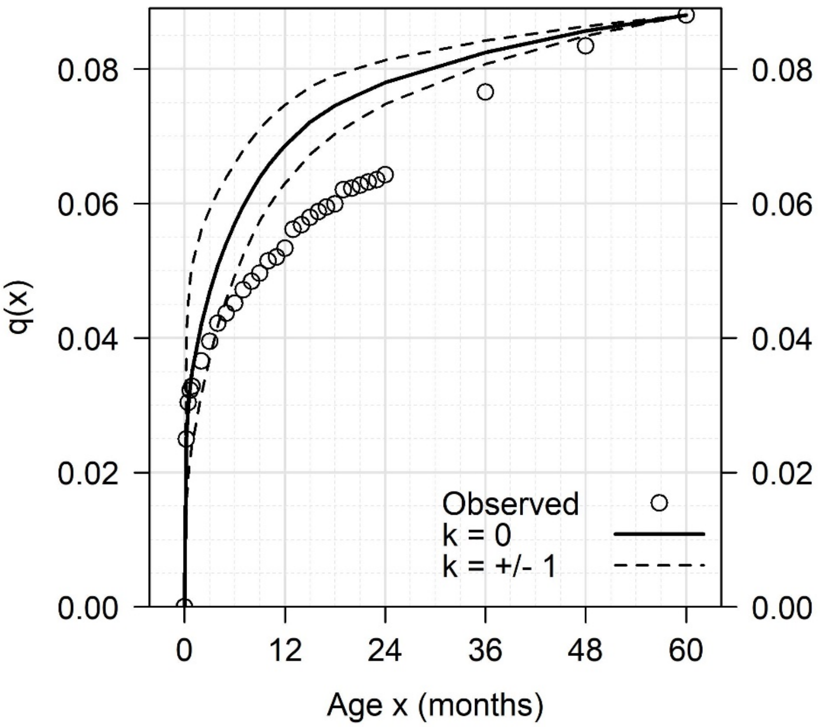
Figure 10: Observed and Predicted values of $_nM_x$ in Jordan, 2015 (both sexes)



Observed = unadjusted VR-based $_nM_x$ values

Predicted = $_nM_x$ values predicted on the basis of the log-quadratic model with the observed VR-based value of q(28d,5y) and k=0 as inputs

Figure 11: Observed values of q(x) in the 2011-12 DHS for Senegal (both sexes), vs. values of q(x) predicted by the log-quadratic model with the same level of q(5y) and k=0±1.



Appendix 1: Best linear estimation of k (analytical solutions)

Given a set of cumulative probabilities of dying $q_i(x)$ and the level of mortality $h_i = \ln[q_i(5y)]$ of a population i:

$$\ln[q_i(x)] = \mathrm{a}_x + \mathrm{b}_x \cdot h_i + \mathrm{c}_x \cdot {h_i}^2 + e_i(x), \qquad \text{(A1)}$$

$e_i(\quad)$ represents the residual at the specific age x, using the coefficients of the model life table $\{\mathrm{a}_x, \mathrm{b}_x, \mathrm{c}_x\}$, as shown by equation A1. These residuals from the average pattern of mortality can be defined as a function of the shape-related coefficient v_x and the scale-parameter k_i, as shown by equation A2:

$$e_i(x) = \mathrm{v}_x \cdot k_i + \epsilon_i(x). \qquad \text{(A2)}$$

The optimal value of k_i minimizes the Mean Squared Error (MSE) of $\epsilon_i(\quad)$, as a weighted function responding to unequal age intervals. In particular, $w(\quad)$ is assumed proportional to the last age interval before the age x, weighting the marginal contribution of an additional equation in the model.

$$MSE_i = \textstyle\sum_{x \in X} w(x) \cdot \epsilon_i(x)^2. \qquad \text{(A3)}$$

As a least squares' solution, the MSE is minimized by making the first derivative of equation A3 with respect to k_i equal to zero, as shown:

$$\frac{\partial MSE_i}{\partial k_i}: -2 \cdot \textstyle\sum_{x \in X} w(x) \cdot e_i(x) \cdot \mathrm{v}_x + 2 \cdot k_i^* \cdot \sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2 = 0.$$

The resulting estimator of k_i is given by equation A4:

$$k_i^* = \frac{\sum_{x \in X} w(x) \cdot e_i(x) \cdot \mathrm{v}_x}{\sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2}. \qquad \text{(A4)}$$

The uncertainty of the model is measured by the variance of the estimator. The first step is to contrast the estimated value of k_i^* with the expected value of k_i, given that:

$$k_i^* = k_i + \frac{\sum_{x \in X} w(x) \cdot \epsilon_i(x) \cdot \mathrm{v}_x}{\sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2}.$$

Hence, the estimated variance of k_i^* is defined as:

$$\mathrm{Var}[k_i^*] = \frac{\mathrm{E}\left[\left(\sum_{x \in X} w(x) \cdot \epsilon_i(x) \cdot \mathrm{v}_x\right)^2\right]}{\left(\sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2\right)^2}.$$

Assuming that prediction errors of different ages are not correlated, the expected value of $\epsilon_i(x) \cdot \epsilon_i(y)$ is zero for each $x \neq y$. Hence, the variance of the estimator can be simplified to be:

$$\mathrm{Var}[k_i^*] = \frac{\sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2 \cdot \mathrm{E}\left[w(x) \cdot \epsilon_i(x)^2\right]}{\left(\sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2\right)^2}.$$

Under the assumption of homoscedastic errors, $\mathrm{E}[w(x) \cdot \epsilon_i(x)^2] = {\sigma_i}^2$ for all x; and the variance of the estimator is a function of the variance of the error of prediction $\epsilon_i(\quad)$, to the form:

$$\mathrm{Var}[k_i^*] = \frac{{\sigma_i}^2}{\sum_{x \in X} w(x) \cdot {\mathrm{v}_x}^2}.$$

The variance of the error of prediction is estimated from the equation A3 increased by a factor $\frac{22}{21}$, considering the 22 ages (or equations) in the model and the degree of freedom lost after estimating k_i^*:

$$\widehat{\sigma_i}^2 = \textstyle\sum_{x \in X} w(x) \cdot \epsilon_i(x)^2,$$

As a result, the variance of k_i^* is given by equation A4:

$$\mathrm{Var}[k_i^*] = \frac{22}{21} \cdot \frac{\sum_{x \in X} w(x) \cdot \epsilon_i(x)^2}{\sum_{x \in X} w(x) \cdot \mathrm{v}_x{}^2}. \qquad \text{(A5)}$$

After some additional steps, the variance of the estimator is redefined as a function of the residuals of the model when the pattern of the mortality is ignored $e_i(x)$ and the optimal value of k_i:

$$\mathrm{Var}[k_i^*] = \frac{22}{21} \cdot \left[\frac{\sum_{x \in X} w(x) \cdot e_i(x)^2}{\sum_{x \in X} w(x) \cdot \mathrm{v}_x{}^2} - k_i^{*2}\right]. \qquad \text{(A6)}$$

Given the standard deviation of k_i^*, 95% confidence intervals where calculated assuming a normal distribution, to the form:

$$k_i^* \pm 1.96 \cdot \mathrm{sd}[k_i^*].$$

Appendix 2: General estimation of h and k, using the method of Lagrange (nonlinear approach for numeric solutions)

The best linear estimation of k_i (application of equation A4) assumes that the level of under-5 mortality (q$_i$(5y)) and at least one other q$_i$(x) value between 0 and 5 are given. However, some applications require a general solution of k_i, when: *i)* q$_i$(5y) is unknown; and/or *ii)* the log-quad model is used for matching/fitting specific functions that are not represented in the estimation, such as mortality rates, durations, and probabilities of dying that do not cumulate from zero (e.g., starting at some point after birth).

Inasmuch as some applications involve a transformation of the log-quadratic model, matching/fitting the log-quad model to some relevant data is a problem of optimization subject to nonlinear constraints. Hence, in the most general case the relevant parameters h_i and k_i would result of solving the problem of constrained optimization through numerical methods. From this perspective, the Lagrangian $\mathcal{L}(\ )$, represents a general problem of matching and optimization, using nonnegative multipliers to add nonlinear constraints to the objective function, to the form:

$$\mathcal{L}(h_i, k_i, \lambda_i) = MSE(h_i, k_i) - \lambda_i \cdot [g(h_i, k_i) - \bar{g}_i], \qquad \text{(A7)}$$

where $MSE(h_i, k_i)$ is the mean squared error of a population i, $g(h_i, k_i)$ is the value to be matched as a function of the parameters of the model, $\bar{g}$ is the numerical value of the constraint, and $\lambda_i > 0$ is the Lagrange multiplier.

The general estimation of the model implies finding the values of (h_i, k_i, λ_i) that will make the partial derivatives of equation A7 equal to zero. Using the Newton-Raphson approach, we multiply the gradient (vector of first derivatives) by the inverse of the Hessian (matrix of second derivatives) to adjust the values of an initial approximation. Assuming this approximation is relatively close to the true solution, the optimal values of (h_i, k_i, λ_i) can be iteratively calculated by equation A8:

$$\begin{bmatrix} h_i^o \\ k_i^o \\ \lambda_i^o \end{bmatrix} = \begin{bmatrix} h_i \\ k_i \\ \lambda_i \end{bmatrix} - \begin{bmatrix} \frac{\partial^2 \mathcal{L}(\ )}{\partial h_i \partial h_i} & \frac{\partial^2 \mathcal{L}(\ )}{\partial h_i \partial k_i} & \frac{\partial^2 \mathcal{L}(\ )}{\partial h_i \partial \lambda_i} \\ \frac{\partial^2 \mathcal{L}(\ )}{\partial k_i \partial h_i} & \frac{\partial^2 \mathcal{L}(\ )}{\partial k_i \partial k_i} & \frac{\partial^2 \mathcal{L}(\ )}{\partial k_i \partial \lambda_i} \\ \frac{\partial^2 \mathcal{L}(\ )}{\partial \lambda_i \partial h_i} & \frac{\partial^2 \mathcal{L}(\ )}{\partial \lambda_i \partial k_i} & \frac{\partial^2 \mathcal{L}(\ )}{\partial \lambda_i \partial \lambda_i} \end{bmatrix}^{-1} \cdot \begin{bmatrix} \frac{\partial \mathcal{L}(\ )}{\partial h_i} \\ \frac{\partial \mathcal{L}(\ )}{\partial k_i} \\ \frac{\partial \mathcal{L}(\ )}{\partial \lambda_i} \end{bmatrix}, \qquad \text{(A8)}$$

given a set of partial derivatives are calculated by:

$$\frac{\partial \mathcal{L}(\ )}{\partial h_i} \approx \frac{\mathcal{L}(h_i + \Delta, k_i, \lambda_i) - \mathcal{L}(h_i - \Delta, k_i, \lambda_i)}{2 \cdot \Delta},$$

and

$$\frac{\partial^2\mathcal{L}(\ )}{\partial h_i \partial k_i} \approx \frac{\mathcal{L}(h_i+\Delta,k_i+\Delta,\lambda_i)-\mathcal{L}(h_i-\Delta,k_i+\Delta,\lambda_i)-\mathcal{L}(h_i+\Delta,k_i-\Delta,\lambda_i)+\mathcal{L}(h_i-\Delta,k_i-\Delta,\lambda_i)}{4\cdot\Delta^2}.$$

Since matching two inputs is also a problem of optimization, equation A7 can be redefined to have two multipliers (one per matching constraint) and no minimization part involved $MSE = 0$, to the form:

$$\mathcal{L}(h_i, k_i, \lambda_{1,i}, \lambda_{2,i}) = -\lambda_{1,i} \cdot [g_1(h_i, k_i) - \bar{g}_{1,i}] - \lambda_{2,i} \cdot [g_2(h_i, k_i) - \bar{g}_{2,i}] \qquad \text{(A9)}$$

Optimal solution of equation A9 is feasible using the same iterative procedure of equation A8. However, the gradient and the Hessian are augmented in one dimension in order to include the partial derivatives of the second multiplier.

Appendix Table A1: List of country-years included in the final Under-5 Mortality Database (U5MD) used for estimating the coefficients of the log-quadratic model

Country	**Years**							**n**
Australia	1921-1971	1973-2014						93
Austria	1970-1994	1996-2016						46
Belgium	1946-1954	1956	1961-1992	2007-2010	2013-2014			48
Canada	1929-1942	1944-1975	1977-1986	1988-1990	1992	1995-1997	1999-2006	71
Chile	1992-2004	2007						14
Denmark	1929-1993	1997	2000-2015					82
Finland	1926-1940	1946-1990	1994	1996-1998	2000-2015			80
France	1953-1966	1975-1992	1996-1999	2001-2015				51
Germany	1991-1994	1996-1997	2001-2007	2010-2015				19
West Germany	1956-1960	1970-1971	1973-1977	1979-1990				24
Ireland	1970-1988	1990-1999	2001-2006	2008-2011	2000-2015			39
Israel	1983-1998	2000-2016						33
Italy	1946-1955	1957-1985	1987-2013					66
Japan	1947-1950	1954-1956	1958-1959	1963	1970-1994	1996-2000	2002-2014	53
Netherlands	1970-1994	1996	1998	2000-2001	2004-2008			34
New Zealand	1970-1971	1973-1975	1977-2013					42
Norway	1936-1992	1995-2001	2003-2012					74
Portugal	1971-1993	1996-1997	2001-2015					40
South Korea	2004-2015							12
Spain	1976-1983	1987-1991	1995-1998	2001-2013				30
Sweden	1934-2002	2004-2012						78
Switzerland	1877-1879	1882-1883	1920-1930	1970-1982	1984-1994	1996	1998-2016	60
UK	1937-1985							49
England and Wales	1982-1991	1993	1996-2001	2005-2012				25
United States	1933-1944	1946-1993	1995-1998	2000-2003	2008-2009	2014-2015		72
Total								**1,235**

Supplementary Materials

Description of the Under-5 Mortality Database (U5MD) and the methodology for calculating harmonized mortality rates by detailed age between 0 and 5 years

Table of Contents

1. Introduction

The Under-Five Mortality Database (U5MD) used in our paper provides detailed distributions of deaths from birth to age 5 by sex. These distributions were retrieved from vital registration records for 25 Western countries and for a time window spreading from the second half of the 19th century to very recent periods.

These Supplementary Materials document: (1) the criteria for selecting countries; (2) the sources of information and the methods used to generate the database; (3) the procedure for harmonizing age intervals; and (4) the methods for estimating mortality indicators.

2. Selection of country-years

We selected countries primarily on the basis of a data quality criterion, with the *Human Mortality Database* (HMD) as a reference. The HMD represents the gold standard for mortality estimates in terms of data quality, and thus we decided to select in our database only country-years available in the HMD. Moreover, the overlap between the HMD and our database allowed us to use some of the relevant HMD information in our estimation procedure (see section 3.2).

One difference, however, is that we did not include countries from the former Eastern Bloc even though they are part of the HMD. Numerous authors have pointed out the departure from international standards for defining and reporting live births and infant deaths in the Soviet Union (Anderson et al. 1994; Anderson & Silver 1986; Davis & Feshbach 1980; Velkoff & Miller 1995). Problems include restricted definitions in terms of weight and gestation time period requirements (and thus undercount of live births), but also misreporting of live births and stillbirths, and underregistration of infant deaths. This has led to significant amounts of understatement in the infant mortality rate within the Soviet Union but also in several European countries from the Eastern Bloc aligned on those definitions and practices (Gourbin & Masuy-Stroobant 1995). Studies have shown that underestimation of infant mortality continued after 1990, including after adopting international standards for the definition of a live birth (Aleshina & Redmond 2005; Guillot et al. 2013; Kingkade & Sawyer 2001). Despite recent improvements in the quality of infant mortality information in the region, we adopted the conservative decision of discarding all the countries of the former Eastern Bloc due to the critical importance of the mortality information at early ages for our model.

We also discarded Greece for similar reasons. According to the HMD report (Agorastakis et al. 2017), the country was affected by significant undercount of neonatal deaths at least until the 1980s. Finally, we removed Iceland and Luxembourg due to small population sizes leading to many zero cell counts in the narrow age intervals used in our database.[1]

As a result of these country exclusions, the U5MD includes 25 countries, instead of 40 for the HMD. The list of countries included in the U5MD is presented in Table SM1.

3. Sources of information

In this section, we describe the sources of information used to build the database. These include two primary sources providing raw death counts: (1) historical demographic yearbooks; and (2)

[1] After 2000, the number of deaths between 0 and 5 years tends to be less than 10 deaths per year in these two countries. By contrast, the average number of deaths tends to be higher than 100 in Scandinavian countries and higher than 1000 in other countries.

the UN repository of vital statistics. These two sources were merged as a single dataset. In addition, we used the HMD as a secondary source to fill potential gaps in the primary sources and supply information about exposure to the risk of death for mortality estimation (i.e., denominators of mortality rates).

3.1 Primary sources of information

We extracted age distributions of deaths from the UN repository of vital statistics and the historical demographic yearbooks. As raw data, death counts are integers. Table SM1 shows the country-years from each of these two sources included in the U5MD. In total, the U5MD contains 1,652 country-years.

3.1.1 Sex

We collected age distributions of deaths for each sex. Country-years for which the breakdown by sex was not available were not included in the database. Results for both sexes combined were obtained by summing sex-specific deaths.

3.1.2 Time periods

Age distributions of deaths are period-specific and annual. (Few exceptions with longer periods are indicated in Table SM1.) The UN repository supplied 734 age distributions from 1970 to 2015. The historical demographic yearbooks provided 1008 age distributions on a broader period going from 1841 to 2001. When both sources provided the same country-year information, we kept the one with more detailed information.

The merged sources have a long time span, covering most of the Western experience of mortality transition, with a decline from levels of under-five mortality from around 400 to less than 5 deaths per 1000 births.

3.1.3 Heterogeneity and harmonization of age intervals

3.1.3.1 Heterogeneity of the length of age intervals

We selected only tabulations providing at least the information for the first month and the rest of the first year. However, in most cases the information has greater age detail. The UN repository covers the first year of life with harmonized age intervals: by days for the first week, by weeks until the 28th day, and by month for the rest of the first year. By contrast, the tabulations of deaths in the historical yearbooks are highly heterogeneous. The tables span the first five years of age unevenly and age intervals were reported through a multiplicity of formats across yearbooks.[2]

For example, in 1905, the yearbook of England and Wales reported neonatal deaths by weeks of age, and postneonatal deaths by months of age. Since 1906, deaths occurring during the first day of life were tabulated separately. More details were introduced in 1931, when deaths occurring in the first week were reported by days of age. (The information also included the number of deaths occurring within the first half hour of life.) However, postneonatal deaths have been grouped in trimesters of age since 1926, with the exception of the period 1952-64 tabulated again by months. Figure SM1 shows the age distribution of deaths for the first year of life as it appears in *The Registrar General's Statistical Review of England and Wales for the year 1946*.

In Belgium and France, for many years in the 19th and 20th centuries, deaths were reported by 5-day age groups for the first 10 or 15 days, and then by 5, 10, or 15-day age groups for the rest

[2] 625 age distributions of deaths cover the first year of life only; 170 the first two years; and 223 the first five years.

of the first month. In both countries, the postneonatal information was tabulated unevenly by months, trimesters or semesters. Figure SM2 shows the age distribution of deaths for the first year of life provided by France's *Statistique annuelle du mouvement de la population* for 1906.

In some cases, few or no details were reported regarding neonatal deaths. For example, from 1890 to 1920, the only detail of Danish yearbooks at neonatal ages was the number of deaths in the first day of life (with the exception of the period 1896-1900 where the yearbooks did not provide any detail). From 1921 onwards, the information for the first week was added. Figure SM3 shows the age distribution of deaths provided for two first years of age by Denmark's *Statistisk Tablevierk* for that year.

There is complete absence of detailed information for the first month of life in Belgium (1841-1861), Italy (1872-1889), New Zealand (1972), Portugal (1940-1954), and The Netherland (1850-1864).

For the postneonatal period, the sources always include at minimum information by trimesters and semesters. After the first year, deaths were mostly reported by single year of age. However, for the second year, the information was tabulated by months in Australia (1921-1924) and by trimesters in Belgium (1841-1861), The Netherlands (1850-1864), Norway (1876-1975), and Sweden (1891-1967).

Both the historical yearbooks and the UN repository have tabulations that include some death counts with unknown age. However, age was always identified at least by single year of age. In total, there were 203 country-years with some unknown-age deaths within single year age groups. Among these, 158 country-years were from Norway. The proportion of deaths with unknown age was less than 1% on average. We redistributed these deaths proportionally across the available detailed age intervals within single-year age groups.

3.1.3.2 Heterogeneity of the format of age intervals

In addition to the diversity of the length of age intervals, the age format also varied across sources. For example, in the UN repository, death counts were uniformly formatted by days with months of 28, 60, 90, etc. days, that is with a year of 360 days. In the historical yearbooks, deaths were reported as integer by "hours'', "days'', "weeks", "months" or "years."

3.1.3.3 Harmonizing age intervals

Against this backdrop of heterogeneity, we harmonized the U5MD in two ways. First, we recoded the original formats of age. In order to estimate the precise exposure time to the risk of mortality, we assumed that the average duration of a year was 365.25 days (considering leap years of 366 days). Therefore, we set the average duration of a month to 30.4375 days (365.25/12). However, when the exact number of days of the first month was available (for example 28 days in Figure SM1 or 30 days in Figure SM2), we kept that exact duration and adjusted the duration of the second month accordingly.

Second, we harmonized the lengths of age intervals by weeks for the first 28 days, by months for the rest of the first year, by trimester for the second year, and by year for the three last years. The interpolation method we used for the harmonization is explained in section 4.

3.1.4 Live births, stillbirths, and false stillbirths

The definition of a live birth has evolved over the 20th century. The early recommendation of the League of Nations in 1925 was the presence of breathing as the vital sign to define a live birth.

From 1950 onwards, the WHO replaced this recommendation by the "any sign of life" criterion, making it more inclusive. This recommendation was progressively but unequally adopted by countries until today.[3] Naturally, the definition of stillbirths varied in concomitance, becoming less inclusive over time.

In addition to signs of life, the League of Nations and WHO also recommended a restricted viability criterion for making the distinction between a stillbirth and a miscarriage. The 1950 WHO definition restricted the viability criteria to a minimum gestational age of 28 weeks or a body length of 35-cm. However, since 1975, the WHO has added and reinforced the use of the birthweight as viability criteria (500g or 1000g for international comparisons). This definition is useful for distinguishing stillbirths from miscarriages but should not have an impact on the estimation of live births and their subsequent mortality.

However, in Belgium and France, a specific definition for stillbirths, adopted under the First French Empire, affected the measurement of mortality for certain time periods. In both countries, live births that had died before civil registration (legally within the first three days after delivery) were registered as stillbirths but tabulated separately from actual stillbirths. These "false stillbirths" were registered in Belgium (by sex) from 1879 to 1955 and from 1958 to 1960 (Glei, Devos, et al. 2017), and in France from 1899 to 1974 (by sex from 1953 to 1974) (Glei, Wilmoth, et al. 2017). However, for these periods, we did not find all the necessary death counts nor all the false stillbirth counts. Among the detailed-age death counts that we found, we only included in the U5MD those for which the sex-specific false stillbirths were also available, that is from 1879 to 1954 in Belgium, and from 1953 to 1966 in France. We added these false stillbirths to the number of registered early-neonatal deaths (first week of age). These false stillbirths are thus taken into account in the estimation of mortality for those periods.

3.2 The Human Mortality Database

We used the *Human Mortality Database* (HMD) as secondary source (Human Mortality Database 2018). The HMD is a public database that provides annual mortality and population data from birth to oldest ages by single year of age, for the purpose of studying human longevity.[4] The HMD provides data for 40 countries. The selection of countries was *"limited by design to populations where death registration and census data are virtually complete, since this type of information is required for the uniform method used to reconstruct historical data series. As a result, the countries and areas included here are relatively wealthy and for the most part highly industrialized".* Our database follows that criterion of virtual completeness by selecting, among the available data in the primary sources of information, only the country-years present in the HMD.

We used death counts from the HMD to fill potential missing information in the vital registration data after the first year of life or later. We also used the HMD exposures to the risk of dying for estimating mortality rates (see section 5).

We thus extracted annual death counts and annual exposures to the risk of dying for each of the first five years of life and for each country-year shown in Table SM1. Note that, for both the deaths

[3] For example, during the Soviet period, countries of the Eastern bloc often included the viability criteria of gestational duration and weight, in addition to signs of life, for identifying live births (Gourbin & Masuy-Stroobant 1995; Guillot et al. 2013).

[4] Therefore, the U5MD can be seen as a complement to the HMD with greater age granularity at the earlier ages of life.

and the exposures, we added up the data of England and Wales, Scotland, and Northern Ireland to obtain the total counts of UK.[5] For consistency, we also aggregated the HMD data for the few periods larger than one year in the U5MD (see Table SM1).

3.2.1 Death counts

We extracted death counts by sex for the first five years of life from the "input data files", that is, the published raw deaths. These data were available in different Lexis areas (squares, triangles, and parallelograms) in the HMD primary sources. Therefore, we adopted the following rules to compute death counts by single year of age and single one-year period (i.e., 1x1 in the HMD terminology):

- When deaths were classified in lower and upper triangles, we summed them up to obtain 1x1 Lexis squares.
- When deaths were classified by cohort in parallelograms centered on exact age ("VV" in HMD notation), we divided them into two triangles under assumption of uniform distribution. We then summed lower and upper triangles to obtain 1x1 Lexis squares.
- Some age intervals were not available by single year in the raw data. In these cases, we used the split age intervals of the HMD "complete data series" (adjusted death counts). More specifically, we used the relative age distributions of those adjusted deaths.
- We excluded data with LDB variable = 0. These correspond to data marked as "not used to create the Lexis database" in the HMD.
- When several death counts exist for Lexis areas within the same 1x1 Lexis square, we summed them up to obtain the full 1x1 count.
- We ignored deaths with unknown age, which in the case of the HMD apply to the entire age range (0 to 100+). We instead used information on deaths with unknown age available in the death tables that we collected for the U5MD and that are specific to the under-five age range.

3.2.2 Exposure to the risk of death

All exposure terms used in our database are taken from the HMD "complete data series" with 1x1 Lexis squares. These exposures are expressed in person-years. In most cases, they correspond to mid-year population estimates derived from census enumerations assuming uniformity in the distribution of events (Wilmoth et al. 2017). However, when data on monthly births were available, the HMD team estimated person-years using this more detailed information instead of making the assumption of uniformity in the distribution of births within a calendar year.

4. Harmonization of age intervals

All mortality estimates were computed for the same harmonized age intervals that are specific to the U5MD. These intervals are weeks for the first 28 days, months for the rest of the first year, trimesters for the second year, and years for the three last years (22 age intervals in total).

The typical approaches for harmonizing age groups rely on interpolation methods. For example, the HMD adopted the cubic spline interpolation of McNeil et al. (1977) applied to cumulative distributions of deaths for splitting nx1 aggregated death counts into 1x1 format within a calendar year. As noted in the HMD methods protocol (Wilmoth et al. 2017), the drawback of the spline

[5] The total count for UK is only available from 1982 onwards. Before, the database provides the counts for England and Wales only (see Table SM1).

approach is that the 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. The decreasing function is generally due to spurious oscillations created by the splines because of strong gradient in the data or non-equidistant points. Negative counts occurred at the oldest ages in the HMD, and in some cases in the U5MD. In the case of the HMD, different constraints were used to address this issue, but they did not produce satisfactory results for the U5MD. We thus adopted an alternative interpolation method based on piecewise cubic interpolation or Hermite-type interpolation (Steffen 1990). This method guarantees a monotonic function in every case.

The method constructs a piecewise cubic interpolation function that passes through N given data points. It uses parabolas to determine the slope of the curve at an interval point *i* passing through points (x_{i-1}, y_{i-1}; x_i, y_i; x_{i+1}, y_{i+1}). Then a piecewise cubic function is constructed for each interval (x_i, x_{i+1}). That way the slope of curve has automatically a continuous first-order derivative over the whole set of points. To ensure the interpolation curves behaves monotonically, the method verifies if the parabolas are monotonic. When it is not the case, the method takes for the slope the smallest of the two secants crossing either (x_{i-1}, y_{i-1}; x_i, y_i) or (x_i, y_i; x_{i+1}, y_{i+1}).[6]

We apply the Steffen's method to cumulative distributions of deaths by sex and for both sexes combined. In order to make the procedure more robust, we only applied the method to country-years that had at least one cut-off point at exact age 7 days or later but before 1 month (28 days or 30 days depending on the data format). We removed 99 country-years at the stage due to the absence of such a cut-off point (already identified in section 3.1.3.1).

5. Mortality estimation

In this section, we explain the methods used for estimating age-specific mortality rates ${}_nM_x$ and cumulative probabilities of dying q(x). Using conventional notations, ${}_nM_x[t, t+1)$ is the mortality rate in the age interval *[x,x+n)* and for the period *[t,t+1)*. *q(x)[t, t+1)* is the probability of dying between age 0 and *x* for the same period. Estimates are period-specific and annual, that is describing the mortality experience of a synthetic cohort in a given year. (As discussed above, there are a few exceptions with estimates pertaining to periods of more than one year.)

We estimated the age-specific mortality rates ${}_nM_x$ by dividing the number of deaths at each age interval ${}_nD_x$ in the time period *[t, t+1)* by the exposure to the risk of dying in person-years ${}_nE_x$ for the same age interval and period, as shown in the following equation:

$${}_nM_x = \frac{{}_nD_x}{{}_nE_x}$$

As explained in section 3.2, we extracted exposures to the risk of dying from the HMD by single year of age ${}_1E_{x'}$ where *x'* is the lower bound of the age interval. In order to estimate the exposure for the subintervals of age of the U5MD (such as weeks and months), we assumed a uniform distribution of exposure by detailed age within each single-year age group. With this

[6] For boundaries, the method uses the same approach but estimating the slope for the first (or last) point with parabolas fitted on the two next (or previous) points, and only using one secant.

assumption, the exposure term is proportional to the length of the age interval *n*, as shown in the below equation:

$$_{n}E_{x}[\mathrm{t},\mathrm{t}+1) = \ _{1}E_{x'}[\mathrm{t},\mathrm{t}+1)\ \cdot\ n$$

where $x' \leq x < x'+1$ and $n < 1$.

We then computed cumulative probabilities of dying *q(x)* under the assumption that mortality rates were constant within each detailed age interval:

$$q(x) = \ 1 - e^{-\sum_{i=1}^{x-1} {}_{n_i}M_i\, n_i}$$

A total of 22 estimates of $_{n}M_{x}$ and q(x) were calculated for each country-year, by sex and for both sexes combined: four estimates by week of age for the neonatal period; 11 estimates by month of age at postneonatal ages; four estimates by trimester of age for second year of life; and three estimates by single year of age for the remaining three years. This gives the following arrays:

For age-specific mortality rates, $_{n}M_{x}$:
$_{7}M_{0(d)}$, $_{7}M_{7(d)}$, $_{7}M_{14(d)}$, $_{7}M_{21(d)}$,
$_{1}M_{1(m)}$, $_{1}M_{2(m)}$, $_{1}M_{3(m)}$, $_{1}M_{4(m)}$, $_{1}M_{5(m)}$, $_{1}M_{6(m)}$, $_{1}M_{7(m)}$, $_{1}M_{8(m)}$, $_{1}M_{9(m)}$, $_{1}M_{10(m)}$, $_{1}M_{11(m)}$,
$_{3}M_{12(m)}$, $_{3}M_{15(m)}$, $_{3}M_{18(m)}$, $_{3}M_{21(m)}$,
$_{1}M_{2(y)}$, $_{1}M_{3(y)}$, $_{1}M_{4(y)}$

For cumulative probabilities of dying from birth to age x, q(x):
q(7d), q(14d), q(21d), q(28d),
q(2m), q(3m), q(4m), q(5m), q(6m), q(7m), q(8m), q(9m), q(10m), q(11m), q(12m),
q(15m), q(18m), q(21m), q(24m)
q(3y), q(4y), q(5y)

6. References

Agorastakis, M., Jdanov, D., & Grigoriev, P. (2017). *About mortality data for Greece*. Retrieved from https://www.mortality.org/

Aleshina, N., & Redmond, G. (2005). How high is infant mortality in central and eastern Europe and the commonwealth of independent states? *Population Studies, 59*(1), 39-54. doi:10.1080/0032472052000332692

Anderson, B. A., Katus, K., & Silver, B. D. (1994). Developments and prospects for population statistics in countries of the former Soviet Union. *Population Index, 60*(1), 4-20. doi:10.2307/3645322

Anderson, B. A., & Silver, B. D. (1986). Infant mortality in the Soviet Union: Regional differences and measurement issues. *Population and Development Review, 12*(4), 705-738. doi:10.2307/1973432

Davis, C., & Feshbach, M. (1980). *Rising infant mortality in the USSR in the 1970's* (Vol. No. 74). Washington, DC: US Bureau of the Census.

Glei, D. A., Devos, I., & Poulain, M. (2017). *About mortality data for Belgium*. Retrieved from https://www.mortality.org/

Glei, D. A., Wilmoth, J., Vallin, M., Vallin, J., & Meslé, F. (2017). *About mortality data for France, total population*. Retrieved from https://www.mortality.org/

Gourbin, C., & Masuy-Stroobant, G. (1995). Registration of vital data: are live births and stillbirths comparable all over Europe? *Bulletin of the World Health Organization, 73*(4), 449-460.

Guillot, M., Lim, S.-j., Torgasheva, L., & Denisenko, M. (2013). Infant mortality in Kyrgyzstan before and after the break-up of the Soviet Union. *Population Studies, 67*(3), 335-352. doi:10.1080/00324728.2013.835859

Human Mortality Database. (2018). University of California, Berkeley (USA) and Max Planck Institute for Demographic Research (Germany), available at www.mortality.org (data downloaded on November 20, 2018).

Kingkade, W. W., & Sawyer, C. C. (2001). *Infant mortality in Eastern Europe and the former Soviet Union before and after the breakup*. Paper presented at the 2001 International Population Conference of the International Union for the Scientific Study of Population, Salvador de Bahia (Brazil).

McNeil, D. R., Trussell, T. J., & Turner, J. C. (1977). Spline interpolation of demographic data. *Demography, 14*(2), 245-252. doi:10.2307/2060581

Steffen, M. (1990). A simple method for monotonic interpolation in one dimension. *Astronomy and Astrophysics, 239*, 443-450.

Velkoff, V. A., & Miller, J. E. (1995). Trends and differentials in infant mortality in the Soviet Union, 1970-90: How much is due to misreporting? *Population Studies, 49*(2), 241-258. doi:10.1080/0032472031000148496

Wilmoth, J. R., Andreev, K., Jdanov, D., Glei, D. A., & Riffe, T. (2017). *Methods protocol for the human mortality database*. Retrieved from https://www.mortality.org/Public/Docs/MethodsProtocol.pdf

Table SM1. Country-years included in the original Under-5 Mortality Database (U5MD)

Country	Years	Sources	
		Statistical Yearbooks	UN database
Australia	1921-71, 1973-2014	51	42
Austria	1970-94, 1996-2016	-	46
Belgium	1841-60, 1861-70*, 1878-84, 1924-54, 1956, 1961-92, 2007-10,2013-14	80	18
Canada	1929-42, 1944-75, 1977-86, 1988-90, 1992, 1995-97, 1999-2006	41	30
Chile	1992-2005, 2007	-	14
Denmark	1890-94, 1896-1993, 1997, 2000-2015	79	41
Finland	1878-1920, 1921-25*, 1926-40, 1946-90, 1994, 1996-98, 2000-15	91	33
France	1855-68, 1877-89, 1891-1947, 1950-66, 1975-92, 1996-99, 2001-15	101	37
Germany	1991-94, 1996-97, 2001-07, 2010-15	-	19
West Germany	1956-60, 1970-71, 1973-77, 1979-90	5	19
Ireland	1970-88, 1990-99, 2001-06, 2008-11	-	39
Israel	1983-98, 2000-16	-	33
Italy	1872-89, 1926-33, 1939-55, 1957-85, 1987-2013	85	14
Japan	1947-50, 1954-56, 1958-59, 1963, 1970-94, 1996-2000, 2002-14	10	43
Netherlands	1850-64, 1970-94, 1996, 1998, 2000-01, 2004-08	15	34
New Zealand	1970-75, 1977-2013	-	43
Norway	1876-1900, 1901-05*, 1906-26, 1927-30*, 1931-92, 1995-2001, 2003-12	93	34
Portugal	1940, 1942-59, 1962, 1970-93, 1996-97, 2001-15	20	41
South Korea	2004-15	-	12
Spain	1976-83, 1987-91, 1995-98, 2001-13	-	30
Sweden	1891-02, 2004-12	111	10
Switzerland	1877-79, 1882-83, 1920-30, 1970-82, 1984-94, 1996, 1998-2016	16	44
UK	1982-91, 1993, 1996-2001, 2005-12	-	25
England & Wales	1905-1985	65	16
United States	1933-44, 1946-93, 1995-98, 2000-03, 2008-09, 2014-15	60	12
Total		923	729

* Deaths aggregated over the indicated range of years

Figure SM1. Age distribution of deaths in The Registrar General's *Statistical Review of England and Wales* for the year 1946

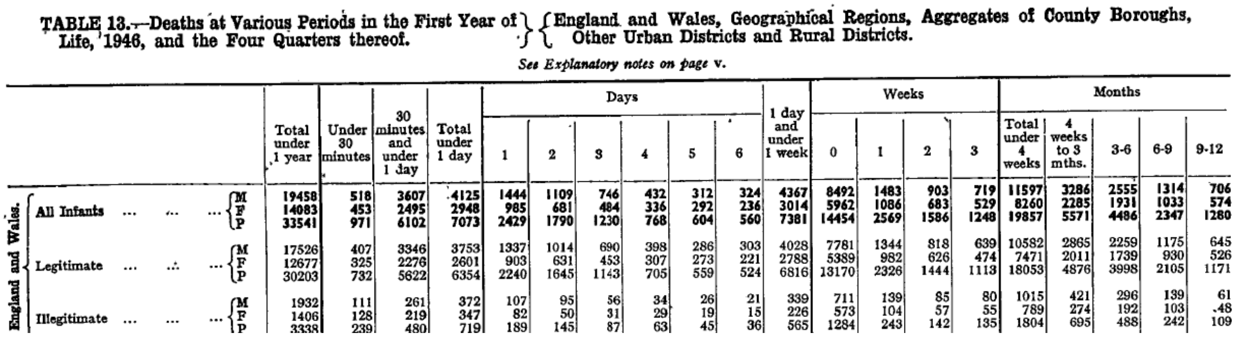
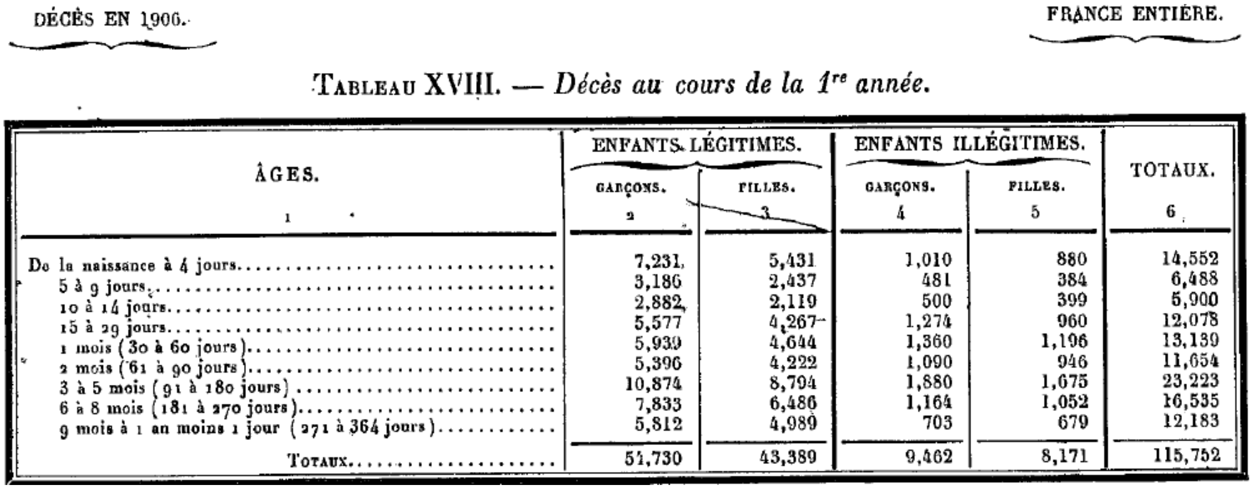


TABLE 13.—Deaths at Various Periods in the First Year of Life, 1946, and the Four Quarters thereof. } { England and Wales, Geographical Regions, Aggregates of County Boroughs, Other Urban Districts and Rural Districts.

See Explanatory notes on page v.

		Total under 1 year	Under 30 minutes	30 minutes and under 1 day	Total under 1 day	Days						1 day and under 1 week	Weeks				Months				
						1	2	3	4	5	6		0	1	2	3	Total under 4 weeks	4 weeks to 3 mths.	3-6	6-9	9-12
England and Wales. All Infants	M	19458	518	3607	4125	1444	1109	746	432	312	324	4367	8492	1483	903	719	11597	3286	2555	1314	706
	F	14083	453	2495	2948	985	681	484	336	292	236	3014	5962	1086	683	529	8260	2285	1931	1033	574
	P	33541	971	6102	7073	2429	1790	1230	768	604	560	7381	14454	2569	1586	1248	19857	5571	4486	2347	1280
Legitimate	M	17526	407	3346	3753	1337	1014	690	398	286	303	4028	7781	1344	818	639	10582	2865	2259	1175	645
	F	12677	325	2276	2601	903	631	453	307	273	221	2788	5389	982	626	474	7471	2011	1739	930	526
	P	30203	732	5622	6354	2240	1645	1143	705	559	524	6816	13170	2326	1444	1113	18053	4876	3998	2105	1171
Illegitimate	M	1932	111	261	372	107	95	56	34	26	21	339	711	139	85	80	1015	421	296	139	61
	F	1406	128	219	347	82	50	31	29	19	15	226	573	104	57	55	789	274	192	103	.48
	P	3338	239	480	719	189	145	87	63	45	36	565	1284	243	142	135	1804	695	488	242	109

Figure SM2. Age distribution of deaths in France's *Statistique annuelle du mouvement de la population* for the year 1906

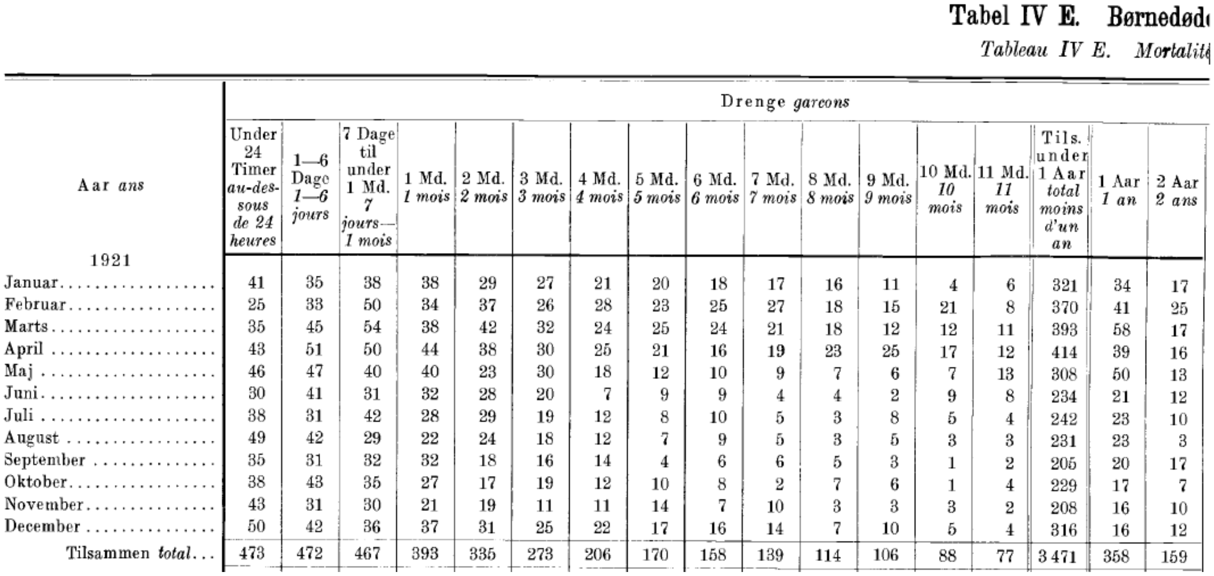


DÉCÈS EN 1906. FRANCE ENTIÈRE.

TABLEAU XVIII. — *Décès au cours de la 1re année.*

ÂGES.	ENFANTS LÉGITIMES.		ENFANTS ILLÉGITIMES.		TOTAUX.
	GARÇONS.	FILLES.	GARÇONS.	FILLES.	
1	2	3	4	5	6
De la naissance à 4 jours	7,231	5,431	1,010	880	14,552
5 à 9 jours	3,186	2,437	481	384	6,488
10 à 14 jours	2,882	2,119	500	399	5,900
15 à 29 jours	5,577	4,267	1,274	960	12,078
1 mois (30 à 60 jours)	5,939	4,644	1,360	1,196	13,139
2 mois (61 à 90 jours)	5,396	4,222	1,090	946	11,654
3 à 5 mois (91 à 180 jours)	10,874	8,794	1,880	1,675	23,223
6 à 8 mois (181 à 270 jours)	7,833	6,486	1,164	1,052	16,535
9 mois à 1 an moins 1 jour (271 à 364 jours)	5,812	4,989	703	679	12,183
TOTAUX	54,730	43,389	9,462	8,171	115,752

Figure SM3. Age distribution of deaths in the Denmark’s *Statistisk Tablevierk* for the year 1921



Tabel IV E. Børnedød

Tableau IV E. Mortalit

Aar *ans*	Drenge *garcons*															
	Under 24 Timer *au-dessous de 24 heures*	1—6 Dage *1—6 jours*	7 Dage til under 1 Md. *7 jours—1 mois*	1 Md. *1 mois*	2 Md. *2 mois*	3 Md. *3 mois*	4 Md. *4 mois*	5 Md. *5 mois*	6 Md. *6 mois*	7 Md. *7 mois*	8 Md. *8 mois*	9 Md. *9 mois*	10 Md. *10 mois*	11 Md. *11 mois*	Tils. under 1 Aar *total moins d'un an*	1 Aar *1 an*	2 Aar *2 ans*
1921																	
Januar	41	35	38	38	29	27	21	20	18	17	16	11	4	6	321	34	17
Februar	25	33	50	34	37	26	28	23	25	27	18	15	21	8	370	41	25
Marts	35	45	54	38	42	32	24	25	24	21	18	12	12	11	393	58	17
April	43	51	50	44	38	30	25	21	16	19	23	25	17	12	414	39	16
Maj	46	47	40	40	23	30	18	12	10	9	7	6	7	13	308	50	13
Juni	30	41	31	32	28	20	7	9	9	4	4	2	9	8	234	21	12
Juli	38	31	42	28	29	19	12	8	10	5	3	8	5	4	242	23	10
August	49	42	29	22	24	18	12	7	9	5	3	5	3	3	231	23	3
September	35	31	32	32	18	16	14	4	6	6	5	3	1	2	205	20	17
Oktober	38	43	35	27	17	19	12	10	8	2	7	6	1	4	229	17	7
November	43	31	30	21	19	11	11	14	7	10	3	3	3	2	208	16	10
December	50	42	36	37	31	25	22	17	16	14	7	10	5	4	316	16	12
Tilsammen *total*	473	472	467	393	335	273	206	170	158	139	114	106	88	77	3 471	358	159