Demography (2022)
DOI 10.1215/00703370-9709538

ADVANCE PUBLICATION

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

Michel Guillot, Julio Romero Prieto, Andrea Verhulst, and Patrick Gerland

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 study 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. It uses a log-quadratic approach in predicting a full mortality schedule between ages 0 and 5 on the basis of only one or two parameters. With its larger number of age-groups, the proposed model offers greater flexibility than existing models in terms of both 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 • 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

ELECTRONIC SUPPLEMENTARY MATERIAL The online version of this article (https://doi.org/10.1215/00703370-9709538) contains supplementary material.

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.

This study proposes 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 in predicting a full mortality schedule between ages 0 and 5 on the basis of only one or two 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 article 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 and 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 (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 offer only 0 versus 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 nonstandard 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 and Woods 1998; Knodel and Kintner 1977; Lantoine and Pressat 1984; Lynch et al. 1998; Manfredini 2004). Second, the Coale and Demeny and the United Nations 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 Murray et al.'s (2003) modified logit system, Wilmoth et al.'s (2012) log-quadratic model, and Clark's (2019) singular value decomposition (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, Murray et al.'s (2003) and Wilmoth et al.'s (2012) models are still constrained by the 0 versus 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 study 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 and (2) explicitly expanding the number of age-groups in the model, especially in the first year of life, thus allowing more flexibility than existing models in terms of both 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 proposed model is based on the Under-5 Mortality Database, a newly compiled database for under-5 mortality by detailed age drawn from high-quality vital registration (VR) data. In its original version, this database contains 1,741 annual distributions of under-5 deaths by detailed age (days, weeks, months, trimesters, and years), representing 25 countries over a time window 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 online Supplementary Materials 1.

Age distributions of deaths were obtained from two primary sources: (1) for historical periods (prior to 1970), these distributions were collected manually from archival sources such as national statistical yearbooks; and (2) for periods from 1970 onward, they were obtained electronically from a data repository compiled by the United Nations Statistical Division.

The original selection of country-years was based on 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 considered only country-years available in the HMD for inclusion in the U5MD. The HMD comprises mostly European countries (31) but also some other industrialized countries (nine). However, we did not include all HMD countries in the U5MD. As discussed in the online Supplementary Materials 1, we excluded countries of the former Eastern bloc because of 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 owing to the small size of the population leading to many zero cell counts in the narrow age-groups that we focus on. This gives us an original database containing 1,741 country-years (see Table 1, column 1).

Starting from this original database, we then removed 276 country-years that did not contain enough information for the full harmonization of the database by age and sex (see Table 1, column 2). Specifically, we removed 81 country-years for which death distributions were not broken down by sex. We also removed 117 country-years in France and Belgium because of insufficient details regarding "false stillbirths," that is, deaths that occurred before corresponding births were registered, which were tabulated separately in these two countries (see online Supplementary Materials 1 for details). Regarding the detail of the age information, the minimum criterion for inclusion in the U5MD was the breakdown of infant deaths in terms of neonatal deaths

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 (1)		Excluded Country-Years Because of Insufficient Information (2)		Excluded Country-Years Because of Reversals in the ${}_7M_{0(d)}$ vs. $q(28d,5y)$ Relationship (3)		Final U5MD for Modeling Purposes (4)	
Country	Year Interval	*n*	Year Interval	*n*	Year Interval	*n*	Year Interval	*n*
Australia	1921–2014	93			1921–1934	14	1935–2014	79
Austria	1970–2016	46					1970–2016	46
Belgium	1841–2014	135	1841–1960	59	1879–1945	28	1946–2014	48
Canada	1929–2006	71					1929–2006	71
Chile	1992–2007	14					1992–2007	14
Denmark	1890–2015	120	1890–1920	30	1921–1927	7	1928–2015	83
Finland	1878–2015	124			1878–1928	47	1929–2015	77
France	1855–2015	142	1855–1952	91			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			1876–1934	52	1935–2012	75
Portugal	1940–2015	61	1940–1954	14	1955–1962	6	1970–2015	41
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	108	1911–1969	48	1877–1883	5[a]	1920–2016	55
United Kingdom	1982–2012	25					1982–2012	25
England and Wales	1905–1985	81			1905–1933	29	1934–1985	52
United States	1933–2015	72					1933–2015	72
Total	1841–2016	1,741	1841–1972	276	1855–1974	246	1920–2016	1,219

[a] Excluded because of their extreme values of $q(5y)$ and the large gap (1884–1919) in the time series. See online Supplementary Materials 2, Figure SM2-1, for details.

(<28 days) versus post-neonatal deaths (28 days–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, trimesters, semesters, and years, and distributed over different age spans (first year of age only versus 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, and 28 days; 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 18, and 21 months; and 2, 3, 4, and 5 years. The harmonization was carried out by interpolating cumulative age distributions of deaths using a spline method developed by Steffen (1990), which ensures that the interpolated curves behave monotonically. We excluded 78 country-years at that stage because of insufficient age details during the first month for performing this interpolation (see online Supplementary Materials 1 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; and (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. 2021).

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 (person-years) 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. Mortality rates for both sexes combined were calculated by aggregating sex-specific deaths and exposures. We then calculated cumulative probabilities of dying $q(x)$ (= $1 - l_x/l_0$ in life table notation) 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. In total, this approach produces a fully harmonized under-5 mortality database for a total of 1,465 country-years, by sex and 22 detailed age-groups (see online Supplementary Materials 1, Table SM1-1, for details).

Evaluation of the Quality of the U5MD

As discussed in the foregoing, 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 (nineteenth century and early twentieth 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 versus a stillbirth. Discussions of international standards for defining live births

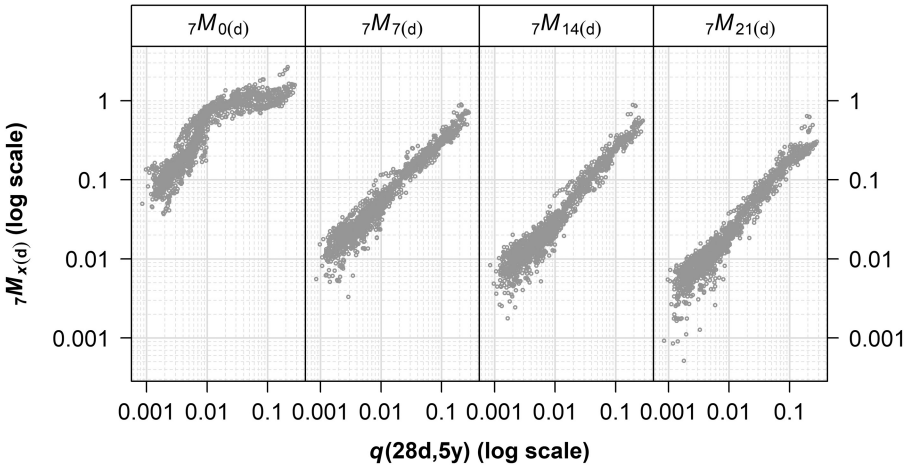

Fig. 1 Relationship between age-specific mortality rates ($_nM_x$) and the probability of dying between age 28 days and 5 years (q(28d,5y)) for each of the first four weeks of life ($_7M_{0(d)}$, $_7M_{7(d)}$, $_7M_{14(d)}$, and $_7M_{21(d)}$) in the harmonized Under-5 Mortality Database (U5MD), for both sexes combined

versus stillbirths started only in the 1920s under the impulse of the League of Nations (United Nations 1954), and distinguishing between live births and stillbirths remains a complex issue even today (Gourbin and 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 1955).

As a result of 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) versus the probability that a 28-day-old child will die prior to reaching the age of 5 years (q(28d,5y)), that is, a mortality indicator not affected by mortality rates for the neonatal period. These relationships are shown in Figure 1.

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, that is, 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 World War I and World War II. 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 ($_7M_{0(d)}$) and mortality between ages 28 days and 5 years are biologically possible. In Switzerland, for example, the level of $_7M_{0(d)}$ 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 $_7M_{0(d)}$ versus q(28d,5y) relationship. Specifically, we examined joint trajectories of the $_7M_{0(d)}$ and q(28d,5y) over time and identified situations in which a local maximum in $_7M_{0(d)}$ was preceded by more than 12 temporally consecutive values of $_7M_{0(d)}$ that were all lower than that local maximum, while no such local maximum was present for q(28d,5y). When such situations were identified, we excluded all years prior to the local maximum in $_7M_{0(d)}$. (When the available time series for a given country started with such a pattern, we removed all points prior to the local maximum in $_7M_{0(d)}$ even if the number of available years prior to that maximum was fewer than 12.) This approach removes the most suspicious patterns while keeping the possibility of a decrease in slope in the $_7M_{0(d)}$ versus q(28d,5y) relationship at higher levels of q(28d,5y). (See online Supplementary Materials 2, Figure SM2-1, for a set of figures showing country-specific time trends in $_7M_{0(d)}$ and q(28d,5y) as well as the relationship between the two indicators, distinguishing years that are excluded based on the foregoing criteria.)

This exclusion criteria removes 241 country-years. We also excluded the first five years of data available for Switzerland (1877–1883), which became isolated, extreme values of under-5 mortality after the removal of the other 241 country-years. In total, 246 country-years were excluded at that stage (see Table 1, column 3). As expected, these country-years pertain mostly to the early years covered by the database: nineteenth century and early twentieth century. (See online Figure SM2-2 for scatterplots distinguishing included versus excluded country-years in the entire database.)

Final Database for Modeling Purposes

The final U5MD that we use for our model includes 1,219 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 1920 until 2016, with levels ranging from around

150 to less than 5 per 1,000. A summary of the available country-years is available in Table 1 (column 4), and full details are provided in Table A1 of the online appendix. The U5MD is freely accessible at https://web.sas.upenn.edu/global-age-patterns-under-five-mortality/.

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 representing the overall *level* of under-5 mortality and the other representing 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:

$$\ln\left[q(x)\right] = a_x + b_x \cdot \ln\left[q(5y)\right] + c_x \cdot \ln\left[q(5y)\right]^2 + v_x \cdot k. \quad (1)$$

As shown in Eq. (1), the model includes a set of age-specific coefficients $\{a_x, b_x, c_x, v_x\}$, whose estimation we describe as follows. When $k=0$, the model predicts a general pattern that 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 nondecreasing 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, ${}_nM_x$, in the left-hand side of Eq. (1). There are four advantages in doing so: (1) the predicted set of $q(x)$ and its corresponding values of ${}_nM_x$ will always agree with the level of $q(5y)$ that is chosen as predictor in the right-hand side of Eq. (1); (2) the model will be more parsimonious, with 21 sets of coefficients versus 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; and (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:

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

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 under-5 age range.

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

The model coefficients in Eq. (1) were estimated in two steps. The first step involved the estimation for each age x of the set of age-specific coefficients $\{a_x, b_x, c_x\}$ regressing $q(x)$ against $q(5y)$ with ordinary least squares. This is shown in Eq. (2), with the subscript i indicating each country-year in our sample of $N = 1{,}219$ observations:

$$ \ln\left[q_i(x)\right] = a_x + b_x \cdot \ln\left[q_i(5y)\right] + c_x \cdot \ln\left[q_i(5y)\right]^2 + e_i(x). \tag{2} $$

The second step uses the age covariance of the residuals $e_i(x)$ in Eq. (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 and 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 an SVD applied to the covariance matrix: $\Psi = \mathrm{V} \cdot \sum \cdot \mathrm{U}$. 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 88% of the total sum of eigenvalues. The R codes that we used to produce model coefficients are available at https://web.sas.upenn.edu/global-age-patterns-under-five-mortality/.

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 0. This is expected given that as x increases, $q(x)$ approaches $q(5y)$.

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

	Females				Males				Both Sexes Combined			
	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	−3.6874	−0.3064	−0.1462	−0.4771	−3.2265	−0.1892	−0.1425	−0.4727	−3.4443	−0.2496	−0.1451	−0.4766
14d	−3.0879	−0.0624	−0.1165	−0.4244	−2.7078	0.0330	−0.1134	−0.4238	−2.8860	−0.0154	−0.1155	−0.4252
21d	−2.6890	0.1033	−0.0968	−0.3918	−2.3603	0.1846	−0.0944	−0.3931	−2.5139	0.1435	−0.0960	−0.3932
28d	−2.4645	0.1925	−0.0864	−0.3673	−2.1500	0.2725	−0.0835	−0.3699	−2.2961	0.2325	−0.0853	−0.3693
2m	−1.9445	0.3793	−0.0653	−0.2860	−1.6300	0.4729	−0.0594	−0.2907	−1.7720	0.4287	−0.0624	−0.2883
3m	−1.7128	0.4418	−0.0591	−0.2310	−1.4171	0.5317	−0.0532	−0.2338	−1.5505	0.4892	−0.0562	−0.2318
4m	−1.5420	0.4857	−0.0551	−0.1926	−1.2680	0.5695	−0.0494	−0.1940	−1.3920	0.5297	−0.0523	−0.1923
5m	−1.3830	0.5344	−0.0501	−0.1663	−1.1330	0.6115	−0.0448	−0.1650	−1.2457	0.5752	−0.0475	−0.1643
6m	−1.2361	0.5824	−0.0453	−0.1461	−1.0026	0.6566	−0.0398	−0.1449	−1.1068	0.6222	−0.0425	−0.1442
7m	−1.1008	0.6282	−0.0406	−0.1311	−0.8833	0.6995	−0.0352	−0.1291	−0.9801	0.6666	−0.0378	−0.1287
8m	−0.9867	0.6671	−0.0367	−0.1190	−0.7805	0.7374	−0.0310	−0.1167	−0.8718	0.7052	−0.0337	−0.1164
9m	−0.8881	0.7011	−0.0332	−0.1079	−0.6904	0.7711	−0.0272	−0.1068	−0.7770	0.7396	−0.0300	−0.1062
10m	−0.7996	0.7325	−0.0299	−0.0998	−0.6133	0.7998	−0.0241	−0.0980	−0.6948	0.7695	−0.0268	−0.0978
11m	−0.7223	0.7603	−0.0269	−0.0923	−0.5478	0.8246	−0.0213	−0.0911	−0.6237	0.7959	−0.0239	−0.0905
12m	−0.6532	0.7854	−0.0243	−0.0863	−0.4867	0.8482	−0.0187	−0.0854	−0.5591	0.8202	−0.0212	−0.0846
15m	−0.4909	0.8439	−0.0181	−0.0710	−0.3465	0.9020	−0.0126	−0.0709	−0.4086	0.8764	−0.0151	−0.0698
18m	−0.3833	0.8835	−0.0136	−0.0601	−0.2600	0.9347	−0.0087	−0.0598	−0.3126	0.9123	−0.0109	−0.0588
21m	−0.3063	0.9119	−0.0103	−0.0514	−0.2031	0.9557	−0.0060	−0.0509	−0.2468	0.9367	−0.0079	−0.0500
2y	−0.2444	0.9335	−0.0078	−0.0433	−0.1565	0.9719	−0.0040	−0.0431	−0.1933	0.9554	−0.0057	−0.0423
3y	−0.1202	0.9696	−0.0037	−0.0207	−0.0660	0.9954	−0.0010	−0.0225	−0.0885	0.9844	−0.0022	−0.0216
4y	−0.0432	0.9912	−0.0011	−0.0085	−0.0266	0.9992	−0.0003	−0.0087	−0.0333	0.9958	−0.0007	−0.0086
5y	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000

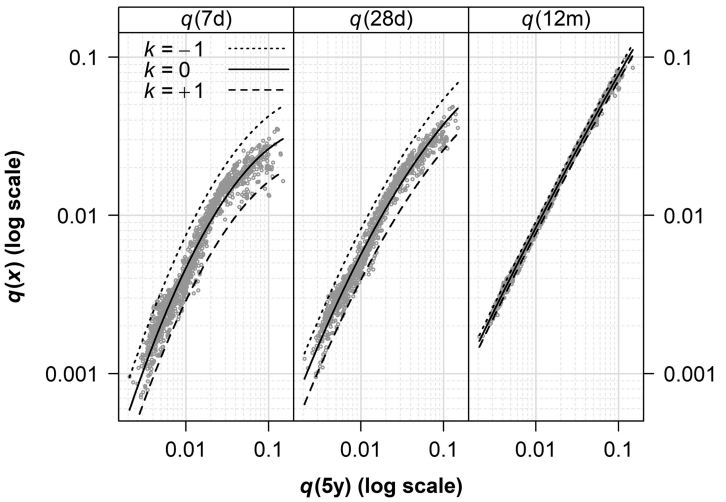

Fig. 2 Relationship between $q(x)$ and $q(5y)$ for x = 7d, 28d, and 12m, with observed values in the final U5MD versus values predicted using the log-quadratic model with $k = 0, -1$, or $+1$, for both sexes combined

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 versus 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 sizable 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 an earlier age pattern of mortality for males.

These features of the model results are illustrated in Figure 2, which shows observed versus predicted values of $q(7d)$, $q(28d)$, and $q(12m)$ when $k = 0$ and when $k = +1$ or -1. Note that almost all data points used for estimating the model are included within this range of values for k. (Country-specific plots for $q(28d)$ are presented in online Supplementary Materials 2, Figure SM 2-3.)

The model results are further illustrated in Figure 3. Panels a and b show how predicted values of $q(x)$ and corresponding values of ${}_nM_x$ 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 1,000, an increasing portion of under-5 mortality takes place below 1 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, that is, "endogenous" causes that have a younger age pattern (Drevenstedt et al. 2008; Galley and Woods 1999; Liu et al. 2012; Rao et al. 2011). Examining the shape of the mortality curves in panel b, 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

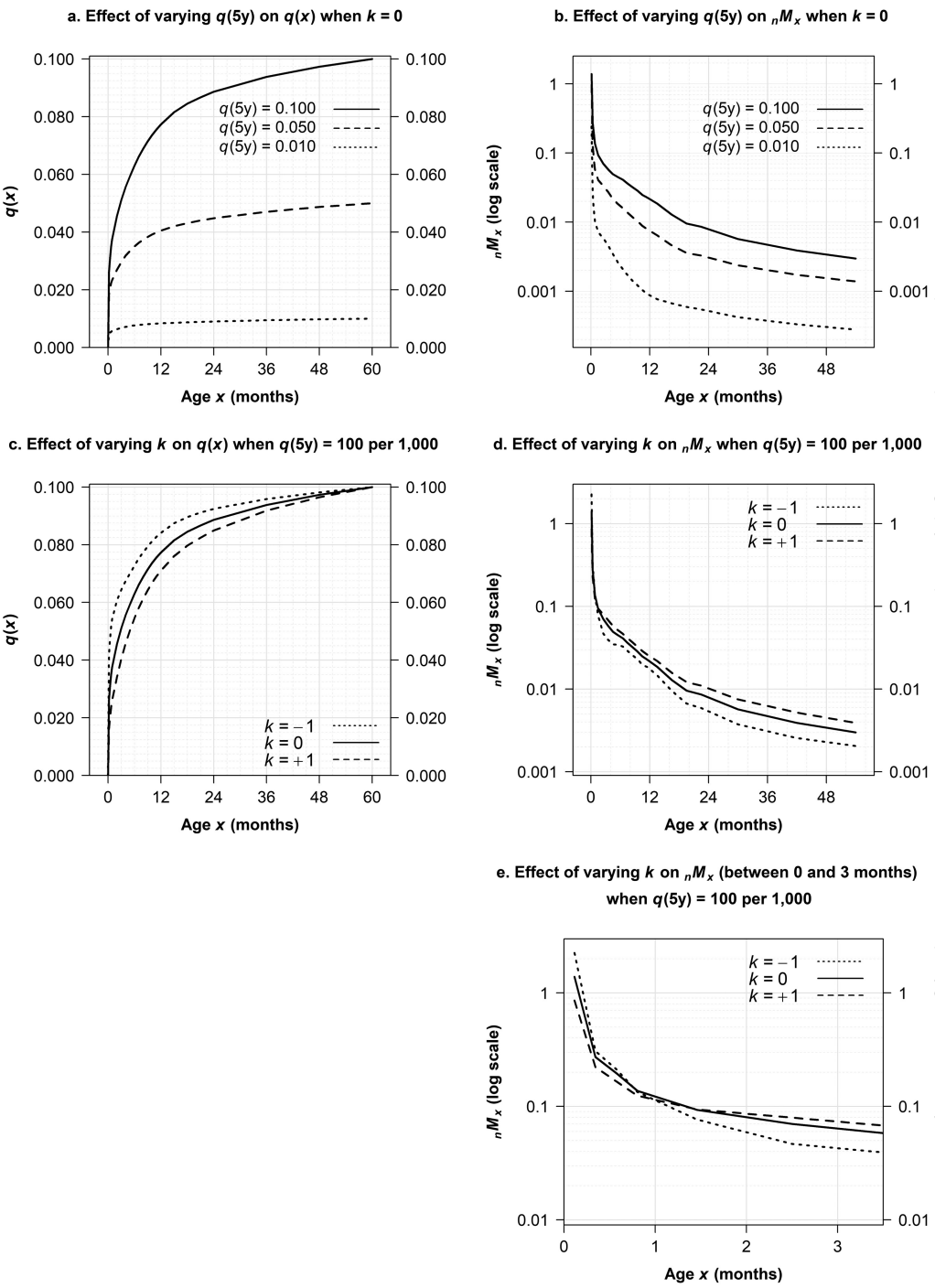

Fig. 3 Effect of varying $q(5y)$ versus k on $q(x)$ and ${}_nM_x$ in the log-quadratic model, for both sexes combined

of $q(5y)$ decreases, the entire mortality curve between 0 and 5 shifts down, with larger relative declines at older versus younger ages.

Panel c of Figure 3 shows the effect of varying k on the $q(x)$ curve at a given level of $q(5y)$ (= 100 per 1,000 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 3 also shows corresponding effects of changing k on ${}_nM_x$ values between 0 and 5 (panel d), with a zoom on the first 3 months (panel e). The mortality curves in this figure all produce the same level of under-5 mortality (100 per 1,000 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 panel e of Figure 3, 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 $q(28d)$ and $q(28d,5y)$. 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 1,000, 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 $q(7d)$ versus $q(7d,5y)$ 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 earlier, 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 that spans between -1.1270 and $+1.5047$, 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 nonmonotonic progression in $q(x)$, which is impossible. As a rule of thumb, users should use the model with k ranging between -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 that, 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 Eqs. (1) and (2):

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

where $e_i(x)$ is the difference between the predicted and observed values of $q_i(x)$ 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 observed value of $q(x)$ in addition to $q(5y)$. For example, knowledge of the infant mortality rate ($q(12m)$) in addition to $q(5y)$ is sufficient for estimating k in a given population.

Alternatively, the value of k for a given population can 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 Eq. (3). Another approach consists of finding the value of k that, 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 for a given population i, 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 Eq. (4) (see the online Appendix 1 for more details):

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

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 data validation purposes.

Figure 4 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 109 per 1,000. 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* (= 0.95 in this example) using Eq. (4). The two entry parameters for Finland in the log-quadratic model are $q(5y) = 0.109$ and $k^* = 0.95$, producing a series of predicted $q(x)$s that fit the observed data remarkably well, with a RMSE of 2.0%. Panel b of Figure 4 also shows how the model fits the observed ${}_nM_x$ series.

The approach discussed in the foregoing 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 5 years. One example of such a 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, owing

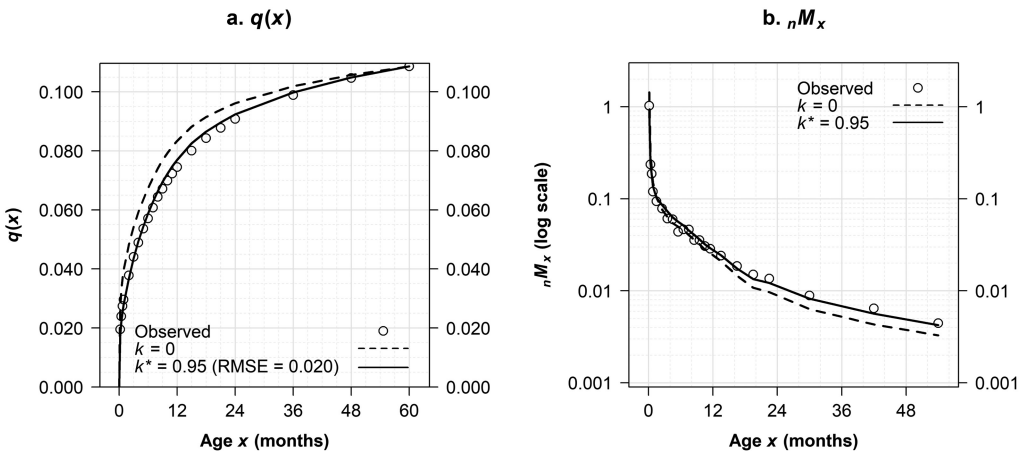

Fig. 4 Observed and predicted values of $q(x)$ and ${}_nM_x$ for Finland, 1933, for both sexes combined

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 value. 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 earlier because of nonlinearities 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 the online Appendix 2. We also use the Lagrange approach in an R package, called "logquad5q0," which we provide as a companion to this article. This package, available at https://github.com/verhulsta/logquad5q0, allows users to use the log-quadratic model to predict a full set of 22 ${}_nM_x$ and $q(x)$ values by detailed age, by sex or for both sexes combined, based on a variety of inputs.

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 0–5 age range will provide a solution. As discussed previously, 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 versus 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 such as

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. This includes, for example, the q(28d,12m) and q(12m,5y) pair, which only covers mortality information above the 28 days threshold, or the q(7d) and q(7d,28d) pair, which only covers mortality information below that threshold. The same conclusion applies when using three 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 with 60% of the country-years for estimating the coefficients of the model $\{a_x, b_x, c_x, v_x\}$ and another with 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, that is, using k^* in Eq. (4).

Table 3 shows the RMSEs for both sexes combined. We report means of the estimates, after preserving the selection 60–40 for a total of 10,000 random samples (without replacement), along with 95% confidence intervals. 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 Eq. (4). The overall adjustment of the model is satisfactory even if a value of $k = 0$ is assumed, with an RMSE of only 4.00%. 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 =1.88% 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, that is, q(28d) and q(12m), respectively. The RMSEs 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. (A version of Table 3 showing mean bias error (MBE) instead of RMSE is presented in the online Appendix Table A2. It shows that the structure of errors is symmetric whenever q(5y) is used as entry point, alone or in combination with other entry points.)

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

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, for both sexes combined

	RMSE for the Following Outcomes:			
Entry Point(s)	All $q(x)$	q(28d)	q(12m)	q(5y)
q(5y) only, $k=0$	0.0400	0.1439	0.0459	0.0000
	[0.038, 0.042]	[0.1365, 0.1514]	[0.0427, 0.0493]	—
q(5y) and q(7d)	0.0242	0.0450	0.0380	0.0000
	[0.0231, 0.0253]	[0.0415, 0.0488]	[0.0360, 0.0400]	—
q(28d)	0.0222	0.0000	0.0372	0.0000
	[0.0210, 0.0234]	—	[0.0348, 0.0396]	—
q(3m)	0.0188	0.0523	0.0299	0.0000
	[0.0179, 0.0199]	[0.0489, 0.0556]	[0.0277, 0.0323]	—
q(6m)	0.0222	0.1081	0.0181	0.0000
	[0.0209, 0.0236]	[0.1008, 0.1165]	[0.0166, 0.0198]	—
q(12m)	0.0324	0.1628	0.0000	0.0000
	[0.0295, 0.036]	[0.1477, 0.1821]	—	—
all $q(x)$[a]	0.0174	0.0558	0.0252	0.0000
	[0.0165, 0.0184]	[0.0524, 0.0594]	[0.0233, 0.0274]	—
q(28d, 5y) only, $k=0$	0.1936	0.3188	0.2051	0.1711
	[0.1795, 0.2086]	[0.2971, 0.3419]	[0.1899, 0.2213]	[0.1584, 0.1845]

Notes: Reported values correspond to the mean of 10,000 random samples; 60% of the life tables were used for estimation and 40% for evaluation. RMSE was calculated from the residuals of 487 life tables (40% of sample). Figures in brackets are 95% confidence intervals.

[a] Using $k=k^*$ (Eq. (4)).

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 versus 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 in 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), RMSEs are substantially higher than when using q(5y) as a predictor (31.88% vs. 14.39%). This is expected given that q(5y) is to a large extent determined by the level of q(28d), making it easier to predict q(28d) on the basis of q(5y) than on the basis of q(28d,5y). RMSEs for other mortality outcomes including q(5y) are substantially lower, because of 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 article. (MBE results presented in the online Appendix Table A2 show that when q(28d,5y) is used as a predictor with $k = 0$, the model has a tendency to produce a downward bias in predicted estimates.

However, this downward bias is small, that is, less than 3% of the true value of q(28d). This slight lack of symmetry is addressed by using nonsymmetrical confidence intervals, as shown in the next section. Note that of all mortality indicators between 28 days and 5 years, q(28d,5y) is the one that produces the smallest prediction 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 28 days, such as q(28d,12m) or q(12m,5y). In that case, prediction errors will be slightly higher: 34.89% with q(28d,12m) and 32.31% with q(12m,5y) versus 31.88% 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 Eq. (4), we obtain in Eq. (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_x, and the optimal value of k_i^* (see the online Appendix 1 for more details):

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

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 $q(x)$ value predicted by the log-quadratic model. This involves calculating predicted values of $q(x)$ in the log-quadratic model using $k_i^* \pm 1.96 \sqrt{\mathrm{Var}[k_i^*]}$.

An illustration of this approach for calculating confidence intervals around predicted $q(x)$ values is provided in Figure 5, using data from Belgium in 1949. We chose this example because of its relatively large remaining prediction errors after estimating k^* (RMSE = 3.4%), making the calculation of confidence intervals particularly relevant. In Figure 5, each predicted $q(x)$ value (panel a) or ${}_nM_x$ value (panel b) is presented with its corresponding 95% confidence interval.

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 1,219 country-years of the final U5MD, the central 95% of the distribution of

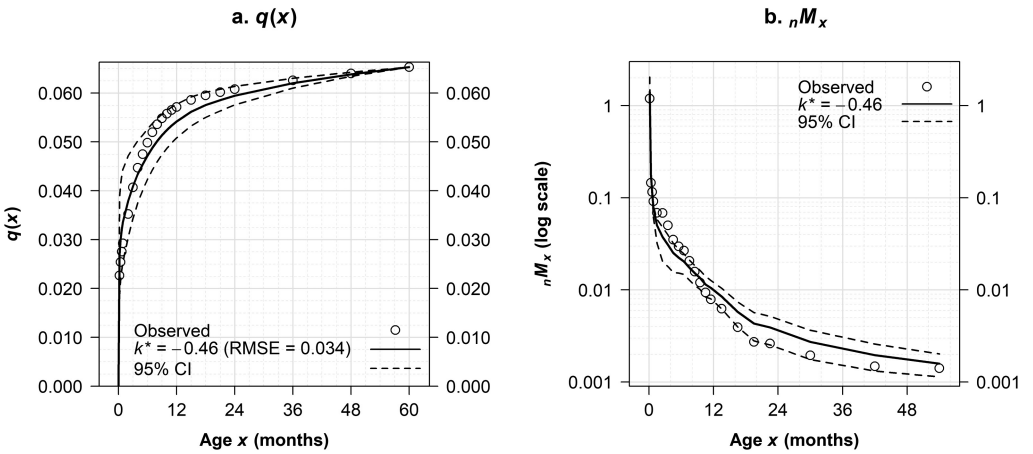

Fig. 5 Observed and predicted values of $q(x)$ and ${}_nM_x$ for Belgium, 1949, for both sexes combined, with 95% confidence intervals

k^* lies between −0.6300 and +0.9314. 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 owing 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-quadratic 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 panels a and b of Figure 6, with data from the 2011–2012 DHS in Honduras. Panel a shows observed $q(x)$ values as well as $q(x)$ values predicted using the log-quadratic model given the observed $q(5y)$ value of 29 per 1,000 and the best-fitting k^* value of 0.06, estimated using Eq. (4). The model fits the $q(x)$ series extremely well, with a RMSE value of 1.7% and narrow confidence intervals. Panel b shows corresponding observed versus predicted ${}_nM_x$ values, illustrating the use of the log-quadratic model for smoothing purposes. The confidence intervals around predicted ${}_nM_x$ values are narrower than suggested by the

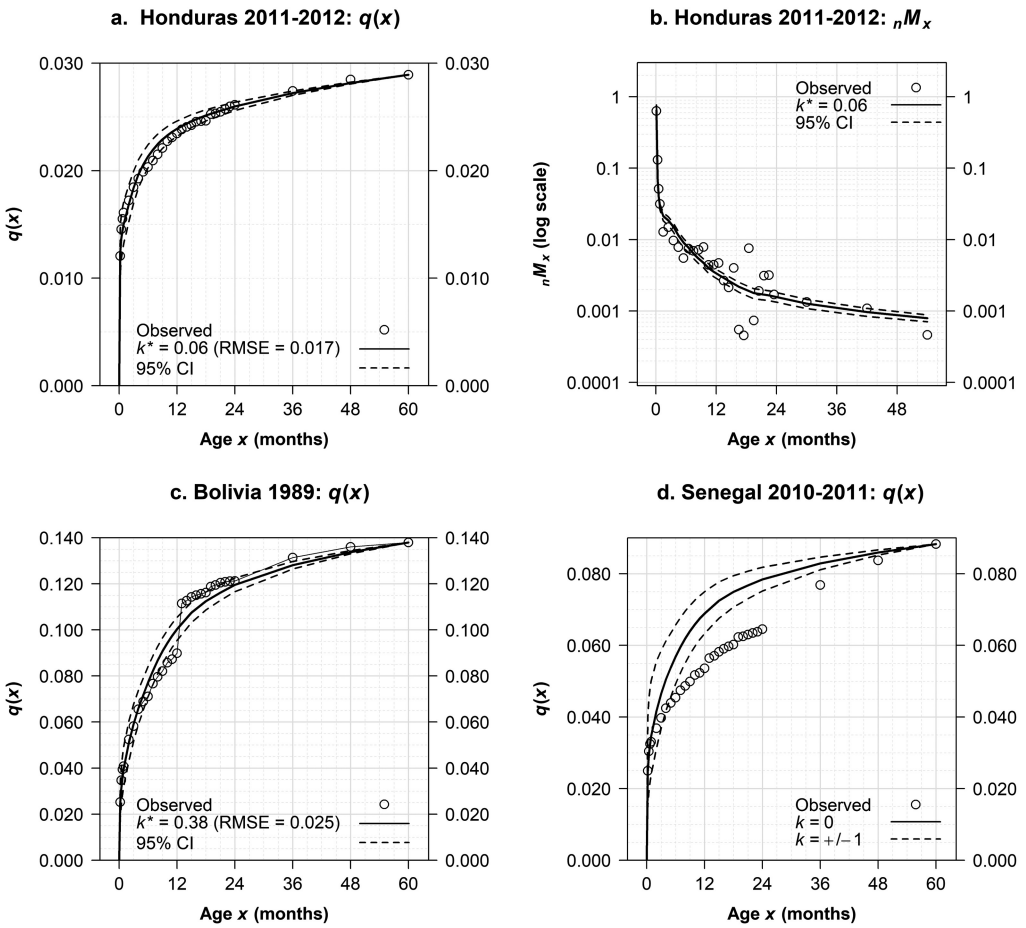

Fig. 6 Observed and predicted values of $q(x)$ and ${}_nM_x$ for selected Demographic and Health Surveys, for both sexes combined. In panel c, k^* is estimated on the basis of observed $q(x)$ points, excluding ages 9 months to 21 months.

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, fixed 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.

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 because of their proximity, for example, 9 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 panel c of Figure 6 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 1,000, but this value is questionable given the presence of such age heaping. This panel also shows predicted values of $q(x)$ with the observed q(5y) value of 138 per 1,000 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 = 2.5%) and predicts an IMR value of 100 per 1,000, that is, 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. (See Romero Prieto et al. (2021) for a more general discussion of this approach.)

In the third application, we show how the model can be used to adjust mortality information in situations where mortality may be underreported at some ages for reasons other than age heaping, for example, owing 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 underreporting. 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 underreporting issues.

We illustrate this type of application using vital registration data from Jordan, a country where VR-based under-5 mortality information appears largely underestimated (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-quadratic model for adjusting under-5 mortality in the country for the year 2015 using the observed value of q(28d,5y) for that year 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 −0.6300 and +0.9314 as discussed in the previous section.

Results of this approach, shown in Figure 7, indicate that the model adjusts the VR estimates of neonatal mortality upward by a factor of more than 2, from 3.7 to 9.9 per 1,000, producing adjusted levels that are consistent with DHS estimates for the same period. Figure 7 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-quadratic model is 17.9 per 1,000, more than 50% higher than the unadjusted level of 11.7 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

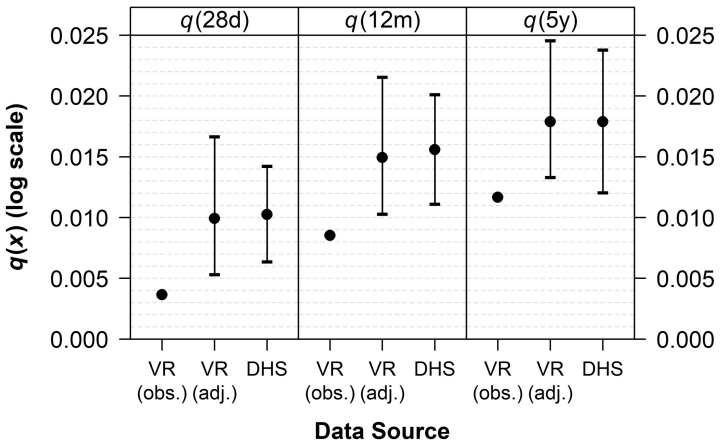

Fig. 7 Levels of neonatal (q(28d)), infant (q(12m)), and under-5 (q(5y)) mortality rates for Jordan in 2015, for both sexes combined, with 95% confidence intervals. VR = vital registration; DHS = 2017–2018 Demographic and Health Survey, with mortality values centered in 2015; obs. = unadjusted VR values; adj. = adjusted VR values using the log-quadratic model with the observed VR-based value of q(28d,5y) and $k = 0$ as inputs.

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 $k = 0$.

In order to further understand the mechanics of this adjustment, we show in online Appendix Figure A1 observed versus 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 onward, the model predicts much higher mortality for the first week. This suggests that underreporting in neonatal mortality in the VR data for Jordan comes primarily from underreporting 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.

Discussion

Mortality between 0 and 5 years 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 and 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 article has a number of strengths.

First, with 22 age-groups between ages 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 versus 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 underreporting 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 Eqs. (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 population 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 geographic 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 and Leridon 1971; Delaunay et al. 2001; Lalou and LeGrand 1996). This unusual age pattern, which has been attributed to a combination of factors, including inadequate weaning foods (Cantrelle and 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 Prieto et al. 2021). As an illustration, we show in panel d of Figure 6 observed $q(x)$ values from the 2010–2011 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 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 data set via coefficients that may then be used for evaluating and correcting data in populations not included in that data set. Nonetheless, Mejía-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 because of 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 panel d of Figure 6 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 the goal to derive a model that can be used as a reference for data evaluation and correction in all low- and middle-income countries requires a thorough evaluation of all the available sources of under-5 mortality information in those countries, an exercise that is beyond the scope of this article. 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 to augment the geographical scope and generalizability of this model.

Conclusion

This article 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. The model uses a log-quadratic approach, predicting a full mortality schedule between ages 0 and 5 on the basis of only one or two parameters.

Results show that our model is able to accurately describe variations in both the level and the 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 geographic scope of the model by gathering the best possible data on under-5 mortality by detailed age in low- and middle-income countries. ■

Acknowledgments Research reported in this article was supported by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health (award R01HD090082). Additional funding was provided by UNICEF. Early versions of this paper were presented at the 2019 PAA meeting in Austin, TX, the Human Mortality Database Symposium in Berlin, the London School of Hygiene and Tropical Medicine, the Johns Hopkins Bloomberg School of Public Health, 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 United Nations Inter-Agency Group for Child Mortality Estimation, 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.

References

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

Agorastakis, M., Jdanov, D., & Grigoriev, P. (2017). *About mortality data for Greece* (Report). Available from https://www.mortality.org/hmd/GRC/InputDB/GRCcom.pdf

Barbieri, M., Wilmoth, J. R., Shkolnikov, V. M., Glei, D., Jasilionis, D., Jdanov, D., . . . Winant, C. (2015). Data resource profile: The Human Mortality Database (HMD). *International Journal of Epidemiology, 44,* 1549–1556.

Bourgeois-Pichat, J. (1951). La mesure de la mortalité infantile: I. Principes et méthodes [The measurement of infant mortality: I. Principles and methods]. *Population (French Edition), 6,* 233–248.

Cantrelle, P., & Leridon, H. (1971). Breast feeding, mortality in childhood and fertility in a rural zone of Senegal. *Population Studies, 25,* 505–533.

Chu, C. Y. C., Chien, H.-K., & Lee, R. D. (2008). Explaining the optimality of u-shaped age-specific mortality. *Theoretical Population Biology, 73,* 171–180.

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,* 1131–1159.

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

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

Croft, T. N., Marshall, A. M. J., Allen, C. K., Arnold, F., Assaf, S., Balian, S., . . . Zweimueller, S. (2018). *Guide to DHS statistics: DHS-7* (Report). Rockville, MD: ICF. 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,* 1286–1295.

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, 105,* 5016–5021.

Galley, C., & Woods, R. (1998). Reflections on the distribution of deaths in the first year of life. *Population, 53,* 921–946.

Galley, C., & Woods, R. (1999). On the distribution of deaths during the first year of life. *Population: An English Selection, 11,* 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)* (Doctoral dissertation). University of Pennsylvania, Philadelphia, PA. Retrieved from https://www.proquest.com/docview/303237054

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,* 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,* e1001299. https://doi.org/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 papers). https://doi.org/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. *Lancet Global Health, 7,* e710–e720. 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,* 391–409.

Lalou, R., & LeGrand, T. K. (1996). Child mortality in towns and villages in the Sahel region. *Population, 51,* 329–351.

Lantoine, C., & Pressat, R. (1984). New aspects of infant-mortality. *Population, 39,* 253–264.

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,* 410–416.

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, 100,* 9637–9642.

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

Liu, L., Johnson, H. L., Cousens, S., Perin, J., Scott, S., Lawn, J. E., . . . Black, R. E. (2012). Global, regional, and national causes of child mortality: An updated systematic analysis for 2010 with time trends since 2000. *Lancet, 379,* 2151–2161.

Lynch, K. A., Greenhouse, J. B., & Brändström, A. (1998). Biometric modeling in the study of infant mortality: Evidence from nineteenth-century Sweden. *Historical Methods, 31,* 53–64.

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

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,* e1002757. https://doi.org/10.1371/journal.pmed.1002757

Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., de Savigny, D., Lozano, R., & Lopez, A. D. (2015). A global assessment of civil registration and vital statistics systems: Monitoring data quality and progress. *Lancet, 386,* 1395–1406. https://doi.org/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,* 165–182.

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,* 66–74.

Romero Prieto, J., Verhulst, A., & Guillot, M. (2021). Estimating the infant mortality rate from DHS full birth histories in the presence of age heaping. *PLoS One, 16,* e0259304. https://doi.org/10.1371/journal.pone.0259304.

Schöley, J. (2019, April). *The age-trajectory of infant mortality in the United States: Parametric models and generative mechanisms*. Paper presented at the 2019 annual meeting of the Population Association of America, Austin, TX.

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

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

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

United Nations. (1982). *Model life tables for developing countries* (Report). New York, NY: United Nations Department of Economic and Social Affairs, Population Division.

United Nations. (1988). MortPak-Lite—The United Nations Software Package for Mortality Measurement: Interactive Software for the IBM-PC and Compatibles [Software]. New York, NY: United Nations Department of International Economic and Social Affairs.

United Nations. (2011). *The Millennium Development Goals report 2011*. New York, NY: United Nations Department of Economic and Social Affairs.

United Nations Inter-agency Group for Child Mortality Estimation (UN IGME). (2019a). *Explanatory notes: Child mortality trend series to 2018* (Report). 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*. Retrieved from https://www.unicef.org/media/60561/file/UN-IGME-child-mortality-report-2019.pdf

Wang, H., Bhutta, Z. A., Coates, M. M., Coggeshall, M., Dandona, L., Diallo, K., . . . Murray, C. J. L. (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. *Lancet, 388,* 1725–1774.

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,* 1–28.

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. (2021). *Methods protocol for the human mortality database* (Report). Retrieved from https://www.mortality.org/Public/Docs/MethodsProtocol.pdf

You, D., Hug, L., Ejdemyr, S., Idele, P., Hogan, D., Mathers, C., . . . Alkema, L. (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. *Lancet, 386,* 2275–2286.

Michel Guillot (corresponding author)
miguillo@sas.upenn.edu

Guillot • Population Studies Center, University of Pennsylvania, Philadelphia, PA, USA; Institut National d'Études Démographiques, Aubervilliers, France; https://orcid.org/0000-0001-5142-6835

Romero Prieto • London School of Hygiene and Tropical Medicine, London, United Kingdom; https://orcid.org/0000-0002-8131-7821

Verhulst • University of Pennsylvania, Philadelphia, PA, USA; https://orcid.org/0000-0003-0571-1330

Gerland • United Nations, Department of Economic and Social Affairs, Population Division, New York, NY, USA; https://orcid.org/0000-0002-2082-3122