DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 53, ARTICLE 26, PAGES 821–896
PUBLISHED 28 OCTOBER 2025
http://www.demographic-research.org/Volumes/Vol53/26/
DOI: 10.4054/DemRes.2025.53.26

Research Article

A parametric survival model for child mortality using complex survey data

Taylor Okonek

Katie Wilson

Jon Wakefield

Contents

Demographic Research: Volume 53, Article 26
Research Article

A parametric survival model for child mortality using complex survey data

Taylor Okonek[1]

Katie Wilson[2]

Jon Wakefield[3]

Abstract

BACKGROUND
Accurate and precise estimates of the under-5 mortality rate (U5MR) are an important summary of the health of a population. Full survival curves on the entire age range are additionally of interest to better understand the pattern of mortality in children under 5. Modern demographic methods for estimating a full mortality schedule for children have been developed for countries with good vital registration and reliable census data but perform poorly in many low- and middle-income countries (LMICs).

OBJECTIVE
In LMICs, the need to utilize nationally representative surveys to estimate U5MR requires additional statistical care to mitigate potential biases in survey data, acknowledge the survey design, and handle aspects of survival data, such as censoring and truncation. We wish to develop parametric and nonparametric approaches for estimating under-5 mortality across time that appropriately utilize complex survey data.

CONTRIBUTION
We propose a parametric approach that is particularly useful in scenarios where data is sparse and estimation may require stronger assumptions. The nonparametric approach we propose provides an aid to model validation. We compare a variety of parametric models to two existing methods for obtaining a full survival curve for children under the age of 5 and argue that a parametric, survey-weighted (pseudo-likelihood) approach is advantageous in LMICs. We apply our proposed approaches to survey data from four LMICs in sub-Saharan Africa. All code for fitting the models described in this paper are available in the `R` package `pssst`.

[1] Mathematics, Statistics, and Computer Science Department, Macalester College, USA. Email: tokonek@macalester.edu.
[2] Department of Biostatistics, University of Washington, USA.
[3] Department of Statistics and Department of Biostatistics, University of Washington, USA.

1. Introduction

Estimates of child mortality rates for specific age groups at a national and subnational level provide important information on the health of a country and inform targeted public health interventions. Historically, estimates of interest have been the neonatal mortality rate (NMR: probability of dying before 1 month of age), the infant mortality rate (IMR: probability of dying before 1 year of age), and the under-5 mortality rate (U5MR: probability of dying before 5 years of age). While these summaries give a rough picture of the pattern of mortality under the age of 5, they do not constitute a complete pattern of mortality before the age of 5. As such, producing a full, continuous survival curve for children under the age of 5 is of interest for informing targeted interventions (Verhulst et al. 2022; Guillot et al. 2022) and quantifying the differences in mortality patterns between countries. Such estimates are particularly important in low- and middle-income countries (LMICs), where rates of child mortality are relatively high. In LMICs, demographic information is primarily collected via nationally representative surveys as opposed to vital registration, and as such, additional statistical care must be taken to adequately account for complex survey designs when computing estimates.

Modern demographic methods for estimating a full mortality schedule for children under the age of 5 have been developed in a high-income country setting where vital registration information is readily available (Guillot et al. 2022; Verhulst et al. 2022). One such method is the log-quad model (Guillot et al. 2022), which uses the recently developed Under-5 Mortality Database (U5MD) (Barbieri et al. 2015) to obtain a continuous curve quantifying the relationship between age and the (log) probability of dying before a given age. This approach uses the Human Mortality Database (HMD, an input to the U5MD) to obtain parameter values, which are plugged into the log-quad model's formula to obtain full, continuous curves. Guillot et al. (2022) note that the patterns of mortality that are estimated from the model are importantly different from the observed data in LMICs. Eilerts et al. (2021), Romero Prieto, Verhulst, and Guillot (2021), and Verhulst et al. (2022) note that sub-Saharan African and South Asian countries typically observe higher levels of the child mortality rate (CMR: the probability of dying between ages 1 and 5 given survival to age 1) for a given IMR when compared to high-income countries. Verhulst et al. (2022) call this a "very late" pattern of under-5 mortality.

Another popular method that makes use of HMD life tables is the singular value decomposition (SVD) approach, described in Clark (2019). Here, the information from HMD life tables are compressed into three or four principal components that summarise observed full mortality schedules over an entire lifetime. Although this approach can be used more generally with other life tables – see Alexander, Zagheni, and Barbieri (2017), for example – the SVD approach used in conjunction with HMD life tables as in Clark (2019) is intended to produce all-age mortality schedules at a yearly scale. As this requires the assumption of a constant mortality hazard within yearly age groups,

this specific application of SVD is not well suited for estimating child mortality, since a constant hazard between ages 0 and 1 year is an unrealistic assumption.

In addition to different patterns of under-5 mortality between LMICs and high-income countries, the difference in data structure (vital registration versus nationally representative surveys) must also be considered when applying or translating demographic and statistical methods across different scenarios. In most high-income countries, vital registration and reliable census information are readily available, hence the mortality data is more granular and potentially subject to fewer biases than are present in data from LMICs. The household surveys that are used to estimate mortality in LMICs typically follow a two-stage, stratified, cluster-sampling design, and are conducted with reasonably high frequency (the aim is every 5 years). There are two major household survey programs collecting data on child mortality: the Demographic Health Surveys (DHS) and the Multiple Indicator Cluster Surveys (MICS). The methods we propose here are applicable to both DHS and MICS, but here we focus on the DHS. To date, DHS has conducted more than 400 surveys in over 90 countries, and is one of the primary data sources used in the production of child mortality estimates by the UN Inter-agency Group for Child Mortality Estimation (IGME) (Alkema and New 2014). Survey-weighted estimates of health outcomes with variance estimates that account for the survey design are preferred when there is enough data to obtain such estimates with high precision, so that resulting estimates are reflective of the underlying population.

As noted in Hill (1995), Lawn et al. (2008), Guillot et al. (2022), and Romero Prieto, Verhulst, and Guillot (2021), surveys such as the DHS may be subject to biases in addition to other data limitations. One example of bias is age heaping, where more children are recorded as having died at particular ages than is truly the case. In DHS surveys, this often occurs at age 12 months (see Appendix A.2 for examples). Additionally, the ages at death of most children are not observed exactly; they are censored. This combined with the need to appropriately account for survey weights and potential biases from age heaping form statistical modeling challenges that are unique to surveys in LMICs; all of these considerations have not yet been addressed simultaneously in the literature in a framework that constructs a full, continuous survival curve.

An additional challenge specific to U5MR estimation is distinguishing between cohort- and period-estimates of mortality. When estimating U5MR, we typically want to obtain period-specific estimates rather than cohort-specific estimates, as the most recent cohort-specific estimates of U5MR we can obtain will always be five years in the past. Period estimates are for "synthetic" children, where the usual approach envisages a hypothetical population of children that live their first five years of life within a single time period. This is in contrast to a real cohort of children who are born in one time period and move through time (periods) as they age. The concept of synthetic people (children or otherwise) allows us to provide estimates of demographic indicators such as life expectancy or U5MR that are a reasonable summary of the current state of the mortality

pattern. In practice, when estimating a demographic indicator for synthetic children, we consider what a real child would contribute to each period as though they were a synthetic child. As detailed in Section 2.3, in a survival analysis framework this corresponds to treating the time period as a time-varying covariate. While existing methods have made use of this approach in a discrete survival setting (Mercer et al. 2015), none have explicitly formulated the problem as that of a time-varying covariate in continuous time.

In this paper we propose a pseudo-likelihood estimate of full mortality schedules for children under the age of 5 in LMICs. Broadly, pseudo-likelihood methods account for complex survey data, both in terms of point and uncertainty estimation. Our approach takes full advantage of the granularity of the data available while accounting for both the survey design and potential biases in the surveys. Rather than assume a model based on data from high-income countries, we instead deal with DHS data directly to obtain an estimate of the survival curve in LMICs at a national level. These methods can flexibly incorporate a variety of parametric distributions, and are readily extendable to subnational estimates. In Section 2 we reframe the production of period estimates for under-5 mortality rates in LMICs using continuous survival models for mortality with a time-varying covariate representation, accounting for potential censoring. In Section 3 we outline our proposed methodology in addition to two existing methodologies currently used in child mortality estimation. Section 4 contains an application to four LMICs, a discussion of model validation, and results. We conclude with a discussion of benefits of our proposed approach, limitations, and future work in Section 5.

2. Survival framework

Framing our estimation problem in a survival context has several benefits, which we describe in greater detail in the following subsections. First, it is straightforward to distinguish between cohort and period estimates by recognizing that the latter creates a left truncation structure in the data. Second, we can directly incorporate ages at death that are not observed at exact times using methods for censored data. Finally, a survival formulation imposes known structure on the resulting estimates – namely, that the cumulative probability of death must be nondecreasing over time, as must be true. As Guillot et al. (2022) note in their paper, the log-quad model can produce estimates that violate this requirement when a certain parameter in their model is estimated outside of an empirical range (see Section 3.3.1). They note that a scenario in which this occurs would be an extrapolation, and in such cases suggest their model should not be used to perform estimation. Our approach provides protection in these situations, which may occur more frequently in LMICs than in high-income countries.

2.1 Time-to-event notation

To begin, we define common demographic and statistical notation that will be used throughout the paper. Mortality, while colloquially referred to as a rate, is typically estimated as a probability that a child dies before a certain age. Let X denote age at death for an individual. We denote the probability that a child died between the ages of x and $x+n$, given that they survived until at least age x, as ${}_nq_x = \Pr(X < x + n \mid X > x)$. With age given in months, a convention we follow throughout, we therefore denote U5MR as ${}_{60}q_0$, IMR as ${}_{12}q_0$, and NMR as ${}_1q_0$.

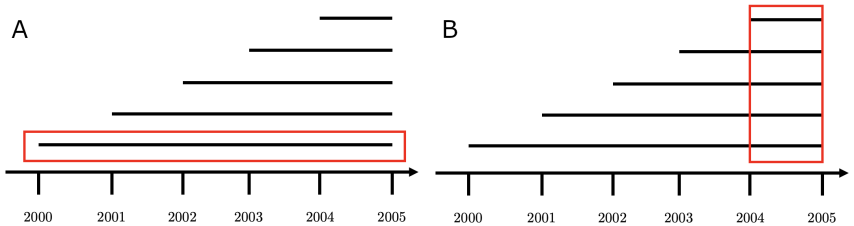
We treat mortality as a time-to-event outcome in a survival framework. In this framework our estimand of interest is the survival curve $S(x)$, which is the probability of surviving to at least age x. We can directly translate ${}_xq_0$ to a survival curve via $S(x) = 1 - {}_xq_0$, and note also that ${}_nq_x = 1 - S(x+n)/S(x)$.

2.2 Period and cohort

An important distinction in demography is period versus cohort estimates. The age-period-cohort distinction is subtle but well documented (see Carstensen (2007), for example) in the demographic literature, but has important statistical consequences, which we elucidate in this section. It is worth noting that if mortality by age is relatively constant across time, cohort and period estimates are roughly equivalent. The need to distinguish between them arises from the fact that they differ when mortality by age changes over time, as we typically observe, particularly in LMICs.

The subset of data used to estimate cohort and period estimates differs. In Figure 1, we illustrate this difference. For simplicity in this explanation, we assume all children are born on January 1 of a given year. We see that the data used to obtain a cohort estimate of U5MR for the cohort born in 2000 consists of children born only in the year 2000. Note that we will always be five years behind schedule in terms of estimate production because we need to observe the full first five years of a cohort before calculating cohort U5MR. The data used to obtain a period estimate of U5MR for the year 2004 contains data from five distinct cohorts: cohorts 2000, 2001, 2002, 2003, and 2004, as seen on the right-hand side of Figure 1. Of note, when obtaining cohort estimates, both age and time align, whereas when obtaining period estimates, age and time are distinct. This is because the age of a synthetic child is not directly tied to time as we observe it. Therefore, we introduce new notation ${}_xq_{n,p} = \Pr(X < x + n \mid X > x, \text{period } p)$ to allow ${}_xq_n$ to vary by period p. For example, the probability a child dies between the ages of 1 and 2 in the year 2001 is ${}_{12}q_{12,2001}$, where the deaths that inform this estimate must come from the cohort of children born in 2000 who survive until at least age 1.

Figure 1: Panel A: Potential lifespans of observations used to obtain a cohort estimate of U5MR for the cohort born in 2000. Panel B: Potential lifespans of observations used to obtain a period estimate of U5MR for the year 2004.



Note: All children are assumed to be born on January 1 of a given year. Horizontal lines indicate the potential lifespans of children up to January 1, 2005.

2.3 Left truncation

It is important to note that when computing period estimates, some of the data will be subject to left truncation. In general, left truncation, also known as late entry, occurs if an individual is not at risk of experiencing the event prior to a certain left truncation time. If not dealt with, left truncation will generally induce selection bias, as individuals who experience the event prior to left truncation time would inherently not be included in the data. Truncation accounts for the potential bias that would be introduced into our estimate from the individuals who were born in the earlier cohorts yet died before our time period of interest. As an example, in Figure 1 (panel B), all individuals born at the beginning of the year 2000 who are still alive by 2004 would be subject to left truncation at age 4 when computing their contribution to the period U5MR estimate for 2004. If we compute the period estimate of ${}_{60}q_{0,2004}$ using five discrete values, ${}_{12}q_{48,2004}$, ${}_{12}q_{36,2004}$, ${}_{12}q_{24,2004}$, ${}_{12}q_{12,2004}$, and ${}_{12}q_{0,2004}$, for each cohort born in 2000 through 2004, respectively, left truncation is dealt with implicitly through the conditional probability structure of ${}_nq_x$, as we will see in the discrete hazards approach (Allison 2014; Mercer et al. 2015) described in Section 3.3.2. This accounts for the artificially smaller risk set that each individual contributes to the period estimate, based on the age at which they enter a given period.

If we are interested in obtaining estimates of U5MR for multiple periods across time, this truncation structure can be incorporated into a model by treating period as a time-varying covariate. This is done implicitly in a discrete hazards approach, but can be done explicitly in the parametric, survey-weighted approach we propose that allows us to use continuous survival models for age. The discretely categorized variable, period, is treated as a covariate that changes through time, and is simply an indicator for which synthetic cohort we are considering.

2.4 Censoring

A final piece of the survival framework is our approach for dealing with censored observations. Not all children die before 5 years of age. Since we are interested in our application in estimating a continuous survival curve for children only up to the age of 5, all children who do not die before 5 years of age will be right censored at that age since they can no longer be at risk of dying under the age of 5 at later ages. Children who die in a time period later than the period in which they were born also contribute right censored observations to those earlier time periods. A survival framework also allows us to deal with interval censoring, where we know only that an event has occurred for an individual between two ages. DHS surveys contain daily observed death dates for children who died before the age of 1 month, monthly, interval censored observations for children who died between 1 month and 24 months (e.g., we may observe a child that dies between the ages of 2 and 3 months), and yearly, interval censored observations for children who died after 2 years of age. There are some rare exceptions in the data where the DHS records more detailed information for particular children. Interval censoring can be appropriately addressed by discretely categorizing observations, as is done in some of the existing approaches described in Section 3, but can also be addressed in the continuous survival framework that we propose.

3. Methods

In Sections 3.1 and 3.2 we describe two proposed approaches for estimating continuous survival curves for synthetic children across multiple time periods. Both methods are based on a finite population, survey-weighted approach – one nonparametric and one parametric – and are novel in the context of child mortality estimation in LMICs. Existing models for comparison are described in Section 3.3.

The statistical methods used in the proposed approaches are often referred to as pseudo-likelihood approaches in the statistical literature (Binder 1983), in which each individual's likelihood contribution is weighted by their sampling weight, and the pseudo-likelihood is maximized to give weighted (pseudo) maximum likelihood estimators (MLEs). The variance of the estimates is computed via sandwich estimation (Binder 1983). In a pseudo-likelihood setting, we are interested in estimating finite population parameters, or summary measures for a population at a fixed point in time, given data from a survey. The practical consequence of this is that the statistical methods used in typical maximum likelihood estimation to quantify uncertainty are inappropriate when we aim to estimate finite population parameters from data with a design that is not simple random sampling. A brief description of the general approach to pseudo-likelihood estimation described in Binder (1983) is given in Appendix B. Historically, survival methods have

been extended to complex survey settings in the context of the Cox proportional hazards model (Binder 1992; Lin 2000; Breslow and Wellner 2007, 2008), but to our knowledge have yet to be extended directly to a setting that involves both left truncation and interval censoring, such as is required in the context of mortality estimation in LMICs.

As noted above, the variance estimation in a pseudo-likelihood approach is what distinguishes this methodology statistically from more commonly used weighted methods. Bootstrap and jackknife procedures have been developed for variance estimation for various complex survey designs, including the two-stage stratified cluster design common to DHS surveys. To obtain bootstrapped variance estimates, $n_h - 1$ clusters are sampled with replacement within strata h, where n_h is the number of clusters in strata h (Rao and Wu 1988). We can then quantify uncertainty via pointwise confidence intervals, constructed using percentiles of the bootstrap samples. A jackknife procedure for the same setting is described in Pedersen and Liu (2012).

3.1 Nonparametric approach

The classic and most popular nonparametric estimate of a survival curve is the Kaplan-Meier estimator (Kaplan and Meier 1958). Let t_i be a time when at least one event (death) occurred, d_i be the number of events that occurred at time t_i, and n_i be the number of children who had not had an event or been censored up to time t_i. Then the Kaplan-Meier estimator of the survival curve at time t is

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right).$$

Under noninformative right censoring and left truncation, the Kaplan-Meier estimator is the nonparametric maximum likelihood estimator (NPMLE) of the survival curve. However, the Kaplan-Meier estimator, in its simplest form, is unsuitable for interval censored data. A generalization of the Kaplan-Meier estimator to arbitrarily truncated and censored observations is the Turnbull estimator (Turnbull 1976). The identifiability of the Turnbull estimator for the interval censoring case we consider was proven by Wang, Gardiner, and Ramamoorthi (1994). In Appendix C.1 we introduce notation for the Turnbull estimator and describe the estimator alongside an example for our motivating application. We develop an extension of the Turnbull estimator that incorporates survey weights in Appendix C.1. Notably, this produces a survey-weighted NPMLE for arbitrarily truncated and censored data, which to our knowledge is the first of its kind.

While this approach can produce point estimates for survival curves in an LMIC context, uncertainty quantification is not straightforward. First, it should be noted that due to the fixed structure of interval censoring present in DHS data (as noted in Section 2.4),

the Turnbull estimator will never, in practice, allow us to obtain any information about the survival curve within age groups defined by this structure. In theory, since deaths are recorded daily in the first month of life and children can die in any given month, given enough data the Turnbull estimator would produce what is essentially a complete survival curve with the only information missing being between 24 hour periods. However, since child deaths are rare, we end up with large gaps of information in the survival curves produced from this method when applied to DHS data.

Second, Groeneboom and Wellner (1992) note that, compared to the Kaplan-Meier estimator, the Turnbull estimator has less appealing asymptotics. In the interval censoring case we consider in this paper, the estimator converges pointwise (i.e., at a fixed value t) at a rate of $(n \log(n))^{1/3}$ to a non-Gaussian distribution. The question of obtaining valid confidence bands for the Turnbull estimator remains an open statistical question. Though some have recommended using a bootstrap procedure for variance estimation (see Sun (2001), for example), the coverage of these procedures is not well justified (and therefore not necessarily correct) due to the rate of convergence and non-Gaussian asymptotics.

Although the bootstrap is not well justified for the Turnbull estimator, we do use the bootstrap procedure appropriate for a two-stage stratified sampling design from Rao and Wu (1988), described at the beginning of this section, to assist with model comparison in our application. In our model comparison approach, we treat the Turnbull estimator as a baseline estimate of the survival curve, and aim to determine whether a given parametric model is 'reasonably' close to the Turnbull estimator. Obtaining some measure of uncertainty for the Turnbull estimator facilitates this comparison.

As a well-justified variance estimator is not available for the Turnbull estimator, we do not recommend using the Turnbull estimator for official estimates of full mortality schedules for children under the age of 5 in LMICs. It is especially important in scenarios where the data does not come from a census or other vital registration source to accurately quantify the uncertainty of estimates. The Turnbull approach does, however, produce a point estimate of the survival function, and therefore is a useful reference when assessing how well a parametric distribution summarizes the pattern of U5MR in LMICs, as its point estimates do not rely on parametric assumptions.

3.2 Parametric approach

In this section, we describe our parametric survival approach if we were interested in estimating child mortality for a single time period. The notation involved in extending the approach to multiple time periods is complex, and is included in Appendix C.2, though we emphasize that the ideas behind the derivation are similar.

Suppose we have children $i = 1, \ldots, n$, whose death times are independent. Let I_i be an indicator that child i's death is interval censored between ages t_{0i} and t_{1i}, where if

$I_i = 0$, child i is either right censored or has an exact death time. Let E_i be an indicator that child i's death is exactly observed at time t_i, where if both E_i and I_i equal zero, t_i is the right censoring time for child i. Finally, let the survey weight for child i be given by w_i.

Following standard probability notation, let $F_{\boldsymbol{\theta}}(x)$ denote the cumulative distribution function for a parametric distribution evaluated at x that depends on a set of unknown parameters $\boldsymbol{\theta}$. Then we can write the pseudo-likelihood as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \left[1 - F_{\boldsymbol{\theta}}(t_i)\right]^{w_i(1-I_i)} \left[F_{\boldsymbol{\theta}}(t_{1i}) - F_{\boldsymbol{\theta}}(t_{0i})\right]^{w_i I_i} \left[f_{\boldsymbol{\theta}}(t_i)\right]^{w_i E_i}. \tag{1}$$

We obtain pseudo-MLEs (Binder 1983) of the distribution-specific parameters by maximizing Equation (1) with respect to the unknown parameters $\boldsymbol{\theta}$. To obtain finite population variance estimates, we use a trick in which we treat our estimator as a weighted total, and use R's `survey` package. The details of this calculation are given in Appendix B. At a high level, the extension of this approach to account for multiple time periods involves rewriting the likelihood in terms of cumulative hazard functions as opposed to cumulative distribution functions, and allowing those cumulative hazards to vary by time period. The resulting pseudo-MLEs have all the convenient statistical properties of MLEs (asymptotic unbiasedness, normality, efficiency) conditional on a correct parametric assumption.

Of note, Schöley (2019) proposes a continuous, parametric approach model for infant mortality. There are similarities between it and our proposed approach, notably the use of a continuous hazard to assist in defining a survival curve for children. The method differs in its focus on the pattern of infant mortality as opposed to U5MR, the use of daily observed deaths from a high-income country that removes the need to account for interval censored observations, and the use of data that does not come from a survey and therefore does not need to account for the survey design. The methods described in Schöley (2019) serve as high income country analogues to our proposed methods, and we consider one family of hazards from Schöley (2019) that produces the best fitting survival curve for US data in our proposed methodology.

Our proposed approach can handle age heaping by lengthening the intervals in which children die, surrounding the time when age heaping is assumed to occur. In our application, we address age heaping at 12 months by interval censoring observations recorded as having died between 6 and 18 months for that entire 12 month period, $[6, 18)$. We chose this window to capture a wide range of potential age heaping surrounding 12 months, but other windows could instead be chosen, depending on assumptions about when age heaping occurs. In aggregating data over these longer intervals, we will lose some precision

in our estimate of the survival curve but should decrease bias due to age heaping, without needing to discard the information that age-heaped individuals provide for our estimates. We emphasize that the benefit of this straightforward approach to addressing age heaping is that the assumptions involved are made transparent – in this case, that the only age heaping in our data occurs between 6 and 18 months. Incorporating additional assumptions about where age heaping occurs is straightforward; we can include additional intervals surrounding the dates where age heaping is thought to occur (for example, ages 3 to 10 days for age heaping at 7 days).

3.3 Existing approaches

As our proposed methodology focuses on providing a continuous, age-specific mortality curve for children under the age of 5, we focus on two existing methods that can provide this, modulo a few assumptions: the log-quad model (Guillot et al. 2022), and the discrete hazards model (Li et al. 2019; Wu et al. 2021). The latter requires the assumption that the discrete hazards ${}_nq_x$, estimated for each x, are constant within the interval $[x, x+n)$ in order to obtain a full survival curve.

3.3.1 Log-quad model

The log-quad model described in Guillot et al. (2022) builds on the approach in Wilmoth et al. (2012), and can provide an estimate of a continuous survival curve from ages 0 to 5 using only an observed or previously estimated ${}_{60}\hat{q}_0$. Other optional inputs to the log-quad model include values ${}_xq_0$ for different ages x. Following Clark (2019), we call the model "empirical" because the coefficients input to the model are not estimated during the modeling process, but instead are computed beforehand using data from the U5MD (Guillot et al. 2022). The model specifies

$$\log({}_xq_0) = a_x + b_x \log({}_{60}\hat{q}_0) + c_x \log({}_{60}\hat{q}_0)^2 + v_x k,$$

where x takes on any one of the 22 values $\{7d, 14d, 21d, 28d, 2m, 3m, 4m, 5m, 6m, 7m, 8m, 9m, 10m, 11m, 12m, 15m, 18m, 21m, 2y, 3y, 4y, 5y\}$. The age-specific coefficients $\{a_x, b_x, c_x, v_x\}$ are provided by the U5MD, ${}_{60}\hat{q}_0$ is input to the model as a fixed covariate, and the parameter k is an optional parameter describing whether the age pattern of mortality is "early" or "late." By "early," we mean that NMR and IMR are higher than what is usually observed when compared to U5MR, and by "late" we mean that NMR and IMR are lower than what is usually observed when compared to U5MR, based on the patterns of mortality before the age of 5 in countries with highly reliable child mortality data, such as those included in the U5MD. When all 22 possible values

for x are supplied to the model, Guillot et al. (2022) propose an uncertainty band around the estimated survival curve, based on the deviation of the shape of the estimated curve from an overall average curve estimated using data from the U5MD. The derivation of their uncertainty band relies on a few key assumptions (including that the deviations of the estimated curve from the overall average curve are independent across ages) that are detailed in Appendix C.4.

Multiple follow-up papers (e.g., Eilerts et al. (2021); Verhulst et al. (2022); Romero Prieto, Verhulst, and Guillot (2021)), as well as Guillot et al. (2022), note that the log-quad model is generally unsuitable for use in LMICs, or in countries with (broadly) early or late patterns of child mortality. This is unsurprising given that the coefficients in the U5MD are estimated from high-income countries, which likely have differing health care systems and structural and programmatic support for decreasing child mortality. Guillot et al. (2022) and Romero Prieto, Verhulst, and Guillot (2021) also note that there are known biases in the data sources available in LMICs. One of these issues, age heaping, can be addressed by excluding data. As we have noted, in DHS surveys especially, age heaping typically occurs at age 12 months. Rather than input all 22 possible age groups into the model for estimating the k parameter, the user may instead leave out a range of ages (Guillot et al. (2022) suggest 9 to 21 months) that they believe covers the ages where data is heaped. Note that this is distinct from treating deaths between the ages of 9 and 21 months as interval censored. The rationale for this approach is that in removing those deaths, the estimated curve will essentially smooth over any age heaping that occurs. It is worth noting that Romero Prieto, Verhulst, and Guillot (2021) find that estimates of IMR from the log-quad model for most surveys/countries they considered did not deviate greatly when compared to a model adjusting for age heaping. Regardless, a significant downside to this approach is that it involves throwing away useful information about the pattern of U5MR, and they additionally caution that their results may not extend to sub-Saharan African or South Asian countries.

While the log-quad model can address age heaping, it has additional characteristics that may be unsuitable in LMICs. Due to its formulation, the log-quad model's prediction of U5MR is identical to the value of U5MR input as a covariate with zero uncertainty (when $x = 5y$, the age-specific coefficients from the model are estimated as $\{a_x, b_x, c_x, v_x\} = \{0, 1, 0, 0\}$). In settings with reliable data, this may be a reasonable (even desirable) property. However, in LMICs where U5MR is often estimated with considerable uncertainty, we do not necessarily want our predicted value of U5MR to align perfectly with a point estimate, but rather to lie within a range of reasonable values defined by the confidence interval for U5MR.

3.3.2 Discrete hazards approach

The discrete hazards approach described in Allison (2014) (as well as Mercer et al. (2015); Li et al. (2019); Wu et al. (2021)) formulates child mortality data in an explicit survival framework. This framework is currently used by the UN and DHS for estimating subnational U5MR in LMICs (Li et al. 2019; Wu et al. 2021). The discrete hazards model splits the time before the age of 60 months into J discrete intervals $[x_1, x_2)$,$[x_2, x_3)$, ..., $[x_J, x_{J+1})$, where $x_{j+1} = x_j + n_j$, $x_1 = 0$. Then, the U5MR can be computed as

$$ {}_{60}q_0 = 1 - \prod_{j=1}^{J}(1 - {}_{n_j}q_{x_j}). \tag{2} $$

Mercer et al. (2015) divide the first 60 months of life for individuals into six intervals, $J = 6$: $[0, 1)$, $[1, 12)$, $[12, 24)$, $[24, 36)$, $[36, 48)$, $[48, 60)$, where $(x_1, \ldots, x_6) = (0, 1, 12, 24, 36, 48)$, $(n_1, \ldots, n_6) = (1, 11, 12, 12, 12, 12)$. Data is tabulated into binomial counts indexed by age group j, and potentially indexed by time period p as well, where the number of observations y_{jp} corresponds to the number of deaths observed in that age group and time period, and the number at risk, defined as N_{jp}, corresponds to the number of children alive in that age group and time period. Note that by construction of the age intervals, we can also estimate NMR and IMR from this model.

Mercer et al. (2015) fit a logistic regression model,

$$ \begin{aligned} y_{jp} \mid N_{jp}, {}_{n_j}q_{x_j,p} &\sim \text{Binomial}(N_{jp}, {}_{n_j}q_{x_j,p}), \\ \text{logit}({}_{n_j}q_{x_j,p}) &= \beta_{jp}, \end{aligned} $$

where β_{jp} is an age-period specific intercept. Pseudo-MLEs of β_{jp} are obtained by fitting this model in R's `survey` package, using the `svyglm()` function. We can use the pseudo-MLEs estimated from the logistic regression model to construct estimates of ${}_{60}q_0$ using Equation (2). Although the binomial likelihood does not reflect the exact data generating mechanism, the correctly specified likelihood closely corresponds to a product of binomial distributions. Many sampling schemes in LMICs (including that used by the DHS) allow data to be aggregated to binomial counts by cluster.

The discrete hazards approach assumes a constant hazard within the specified age groups. Therefore, while we can estimate a full survival curve for children under the age of 5, we know its shape will not be realistic, as the probability of survival should change smoothly with age rather than make discrete jumps. To obtain a more continuous survival curve, we could have 60 age groups for each 1-month breakdown in the discrete hazards approach, if the data were available at a monthly level for all 60 age groups. There

is a balance here between flexibility and parsimony: The model fitted with more age groups better reflects the underlying smooth changes in hazard, but each hazard estimate is less precise than we might get fitting a more parsimonious model (if that model is appropriate).

In contrast with the log-quad model, age heaping can be handled in the discrete hazards model by construction of the age intervals. For example, one could consider age intervals $(J = 7)$: $[0, 1)$, $[1, 9)$, $[9, 21)$, $[21, 24)$, $[24, 36)$, $[36, 48)$, $[48, 60)$, where we group deaths recorded between the ages of 9 and 21 months into a single age group. Additional notes on the discrete hazards model in conjunction with DHS surveys are in Appendix C.5.

4. Application

We apply our proposed parametric pseudo-likelihood approach to child mortality data from Burkina Faso, Malawi, Senegal, and Namibia. We chose single DHS surveys from each of these countries, and used the proposed approach to obtain continuous survival curves for the time periods $[2000, 2005)$ and $[2005, 2010)$ to demonstrate the ability of our approach to produce period estimates throughout time. The data used in the application is described in detail in Section 4.1, and all parametric models considered are catalogued in Section 4.2. We additionally fit a survey-weighted version of the Turnbull estimator, with boostrapped confidence bands, to informally compare the parametric approaches, as described in Appendix C.1.

For further comparison, we contrast our approach to estimates from the log-quad model using all 22 age inputs (calculated from the Turnbull estimate) and the discrete hazards model from Mercer et al. (2015).

For all parametric approaches, we estimate the survival curves in each time period, uncertainty bands surrounding each survival curve (95% confidence bands based on finite population variances for all approaches other than log-quad, and the derived uncertainty band from Guillot et al. (2022) for the log-quad approach), and estimates of NMR, IMR, and U5MR from these survival curves. We note that the uncertainty band for the log-quad model does not have a clear interpretation as a 95% confidence interval for the mortality rate at a given age, and point readers to the derivation in the Supplement of Guillot et al. (2022) for details.

Software for implementing the proposed methodology is available in the `R` package `pssst`, available at https://github.com/taylorokonek/pssst.

4.1 Data

All data used in our application comes from the DHS program. Child death data is collected via interviewing mothers and asking them the birth and death dates of all children they have had. We treat deaths prior to one month as exact and interval censored afterwards with the interval given as a single month or a single year depending on when the child died, as discussed previously in Section 2.4.

It has previously been noted that DHS surveys are subject to potential biases that may negatively impact the resulting estimates of child mortality (Hill 1995; Lawn et al. 2008; Guillot et al. 2022). The main concern for estimates of mortality under the age of 5 years is age heaping at age 12 months, where more children are recorded as having died at 12 months than would otherwise be expected. Lawn et al. (2008) additionally note that age-heaping in DHS surveys may occur at 7 days, 14 days, and 1 month. As noted in Section 3.2, in our proposed approach we address age heaping at 12 months by interval censoring all observations recorded as having died between 6 and 18 months for that entire 12 month period $[6, 18)$. Additionally, we compare estimates from our proposed approach that account for age heaping in this way to estimates that do not make this adjustment. Further details relating to DHS survey design can be found in Appendix A.1.

4.2 Parametric models

The parametric distributions considered for our proposed approach are listed in Table 1. The exponentially truncated shifted-power (ETSP) family of hazards we consider is slightly different than that considered in Schöley (2019), as we set $c = 0$ as opposed to estimating it via profile likelihood. This parameter c can be fixed in our applications, as the finest time scale we have in our observations is daily, and the c parameter controls the hazard in the first hours of life. The generalized Gamma distribution is parametrized as in the `flexsurv` package in `R`, as it is more numerically stable than the original parameterization (Prentice 1974).

Table 1: Parametric distributions considered and their characterizations in terms of a probability density function $f(x)$ or hazard $h(x)$

Distribution	Characterization	Parameters
Piecewise Exponential	$h(x) = \beta_0 I[x < 1] + \beta_1 I[1 \leq x < 12] + \beta_2 I[x \geq 12]$	$\beta_0, \beta_1, \beta_2$
Gompertz	$f(x) = \beta k e^{k+\beta x - k e^{\beta x}}$	β, k
Weibull	$f(x) = \beta k (\beta x)^{k-1} e^{-(\beta x)^k}$	β, k
Lognormal	$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(\log(x)-\mu)^2}$	σ, μ
Generalized Gamma	$f(x) = \frac{\lvert Q\rvert (Q^{-2})^{Q^{-2}}}{\sigma x \Gamma(Q^{-2})} e^{Q^{-2}(Q\omega - e^{Q\omega})}$	Q, σ, ω
Exponentially truncated shifted power (ETSP)*	$h(x) = a(x+c)^{-p} e^{-bx}$	a, b, c, p

Note: *The ETSP hazard as described in Schöley (2019) contains four parameters, but in our applications we set $c = 0$.

4.3 Model validation

To assist with model validation, we fit a survey-weighted version of the Turnbull estimator, with bootstrapped CIs, to provide a guideline for how well each of the parametric distributions is able to capture the underlying survival curve in each time period. This is treated as a reasonable reference point for the underlying survival curve as it is free of parametric assumptions. However, despite our use of bootstrapped CIs there are no well-justified variance estimates for the Turnbull estimator (Section 3.1), making our comparisons to the Turnbull estimator only crudely calibrated.

Let a sample k at age x from the boostrapped distribution of the Turnbull estimate at age x be denoted $\tilde{\theta}_x^{(k)}$, and a sample k from the asymptotic distribution of the parametric survival curve at age x be denoted $\hat{\theta}_x^{(k)}$. We obtain $k = 1, \ldots, 500$ samples, and compute $\hat{\theta}_x^{(k)} - \tilde{\theta}_x^{(k)}$ to obtain samples from the empirical distribution of the difference between the Turnbull and parametric distribution at a given age x.

We calculate the proportion of uncertainty intervals derived from $\hat{\theta}_x^{(k)} - \tilde{\theta}_x^{(k)}$ at ages x that contain 0 as a rough estimate of how closely each parametric model aligns with the Turnbull estimate. This is not a formal hypothesis test but rather a means of assessing how close the parametric estimate is to the Turnbull estimate while accounting for uncertainty in both estimates.

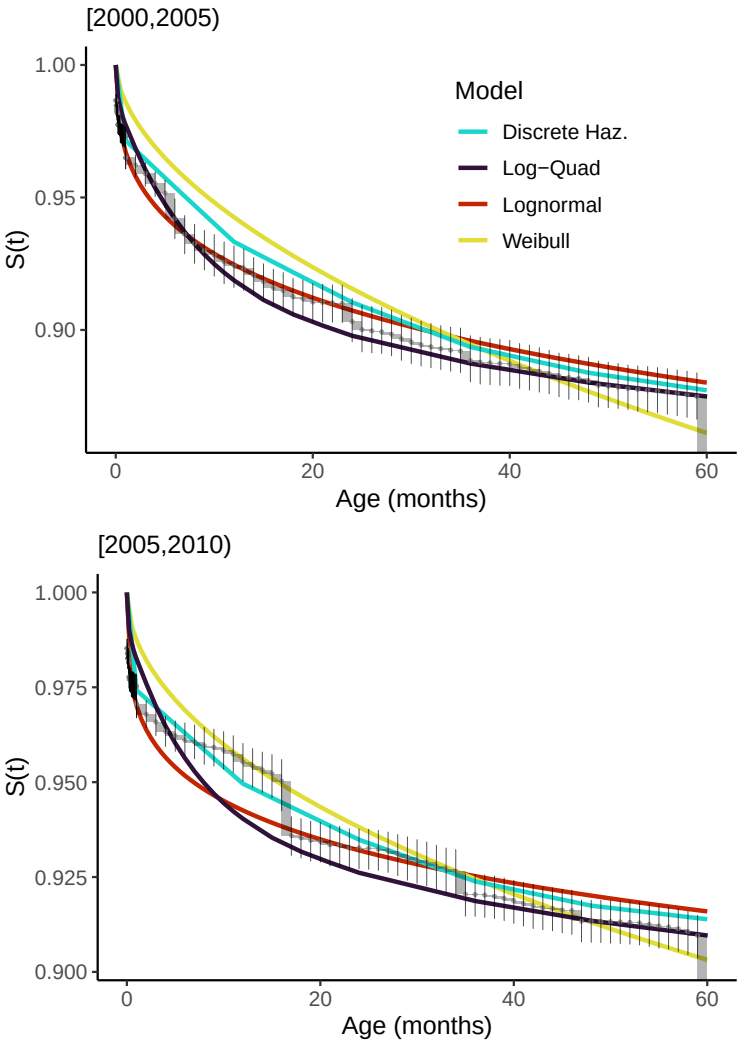
4.4 Results

In this section, we display a subset of results from the application of the seven parametric models, log-quad model, and discrete hazards model to DHS data from Malawi. The results shown for Malawi are representative of the results for Burkina Faso, Senegal, and Namibia, which can be found in the Appendices. Additional results for all four countries can be found in Appendix D, with comparisons to models where the data is not adjusted for age heaping in Appendix E. Results for models that adjusted for age heaping were very similar to results for models that did not adjust for age heaping.

4.4.1 Proposed approach

In Figure 2 we display the fitted survival curves for the Weibull and lognormal models using our proposed methodology in both time periods for Malawi, and compare them to the Turnbull estimator, log-quad model, and discrete hazards model. While other parametric models were estimated in addition to the Weibull and lognormal, we chose to include only these two in Figure 2 for visual clarity (two parametric estimates as opposed to six). The Weibull was chosen because it is one of the most commonly used parametric survival models, and the lognormal was chosen because the fit was particularly close to the Turnbull estimator. Additional visualizations for all parametric models fit can be found in Appendix D.

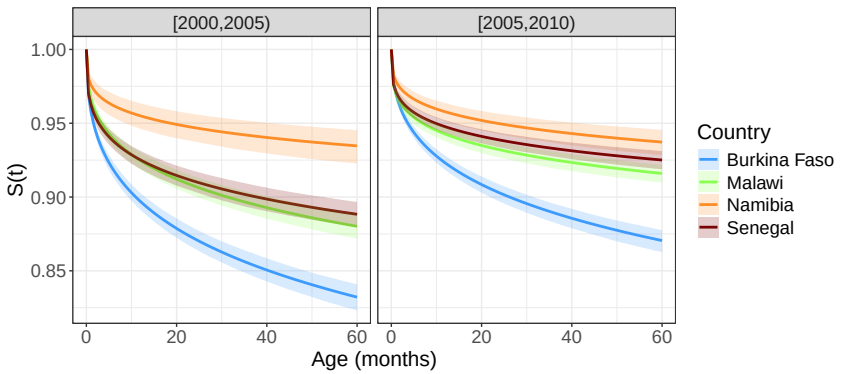
Figure 2: **Estimated Weibull and lognormal survival curves for time periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) for Malawi, compared to estimated survival curves from the discrete hazards and log quad approach**



Note: The Turnbull estimator is displayed as the gray step function, uncertainty boxes, and uncertainty intervals.

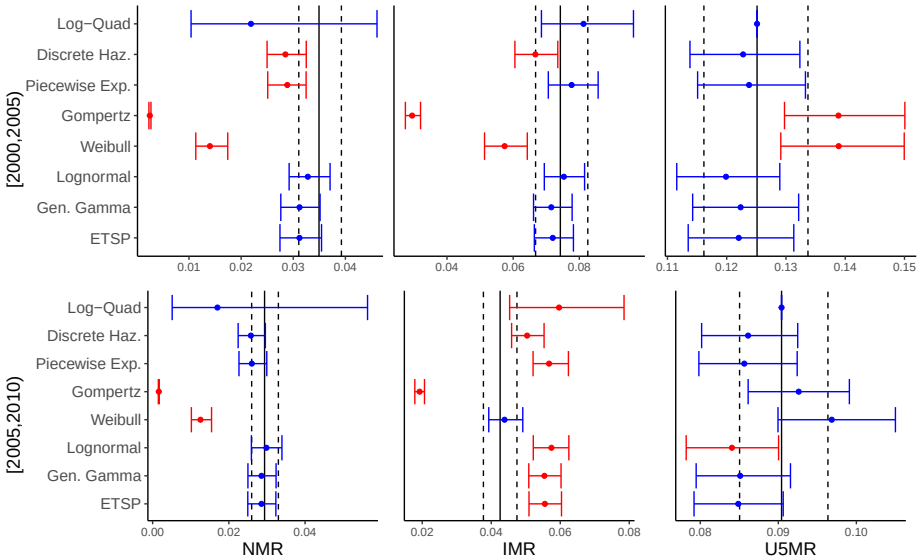
Compared to the Turnbull estimator, the Weibull model tends to estimate higher survivorship at younger ages and lower survivorship at older ages. The lognormal model captures the sharp increase in mortality within the first 12 months of life more accurately than the Weibull model. In Figure 3 we compare estimated lognormal survival curves across all countries in our application and both time periods.

Figure 3: Estimated lognormal survival curves for time periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right) for Burkina Faso, Malawi, Namibia, and Senegal



Showing just summary measures of mortality (NMR, IMR, U5MR), we see the same patterns in Figure 4. The Weibull model in each time period underestimates NMR and overestimates U5MR relative to the Turnbull estimator, particularly for the period $[2000, 2005)$. In contrast, the lognormal model confidence intervals cover NMR, IMR, and U5MR in both time periods, with the exception of IMR in $[2005, 2010)$, where only the Weibull model captures the Turnbull estimate, and U5MR in $[2005, 2010)$, where the lognormal confidence interval is slightly too low to capture the Turnbull estimate of U5MR. As seen in Table 2, the intervals for the estimated difference between the Weibull estimates and Turnbull estimates capture zero for 44% and 60% of monthly ages prior to age 60 months for $[2000, 2005)$ and $[2005, 2010)$, respectively. In contrast, the intervals for the estimated difference between the lognormal estimates and Turnbull estimates capture zero for 93% and 80% of monthly ages prior to age 60 months. This aligns with the visualizations to suggest that the lognormal model is a better parametric fit for the mortality curve for children under the age of 5 than the Weibull model. The log-quad model is not included in Table 2 because it is not a model you can sample from based on the way the uncertainty bands are defined, and therefore our model validation approach cannot apply.

Figure 4: Estimates of NMR, IMR, and U5MR for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)



Notes: Turnbull point estimates are denoted by vertical black lines, with dashed vertical black lines denoting 95% confidence intervals. Horizontal error bars are blue if the interval captures the monthly discrete hazards point estimate, or red if the interval does not capture the monthly discrete hazards point estimate. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model, where uncertainty is calculated as in Guillot et al. (2022).

Table 2: Model validation results

Country	Period	Discrete Hazards	Piecewise Exponential	Gompertz	Weibull	Lognormal	Generalized Gamma	ETSP
Burkina Faso	(2000, 2005)	48	40	12	17	52	**93**	33
	(2005, 2010)	60	48	9	14	65	70	57
Malawi	(2000, 2005)	**73**	**76**	18	44	**93**	**94**	**94**
	(2005, 2010)	**73**	61	12	60	**80**	**82**	**81**
Senegal	(2000, 2005)	**72**	**73**	19	28	**86**	**84**	**85**
	(2005, 2010)	**72**	**73**	13	38	**85**	63	**87**
Namibia	(2000, 2005)	**86**	**86**	25	56	**100**	**100**	**99**
	(2005, 2010)	**85**	**86**	29	53	**100**	**100**	**99**
Total	All	**71**	68	17	39	**83**	**86**	**79**

Notes: Percentage of samples (out of 500) from $\hat{\theta} - \tilde{\theta}$ that contain 0 for all parametric models, countries, and periods. Results that contain more than 70% of samples noted in bold. The final row contains the overall percentage across all countries and periods for a given model.

4.4.2 Comparison to existing approaches

The log-quad approach performs adequately, though it is important to note that while the point estimates may be reasonable, the uncertainty quantification is less so. In particular, U5MR is assumed to be estimated with no uncertainty. This is not a desirable property of this approach since our estimates of U5MR that are input to the log-quad model are themselves estimated with uncertainty. Second, we note that the uncertainty bands around the log-quad point estimates are in general much wider than the confidence bands for the parametric models. The confidence bands surrounding the parametric models may be interpreted at each age x with a 95% confidence interval interpretation based on resampling observations from the finite population, whereas the uncertainty surrounding the log-quad model does not have as straightforward of an interpretation. Furthermore, out of all the analyses conducted, only the log-quad models for Namibia (both time periods) provided estimates and confidence bands that would be considered reasonable by Guillot et al. (2022). All other countries either had estimated values for certain parameters outside the range suggested by Guillot et al. (2022) or an increasing hazard by age in the uncertainty interval computed, which is unrealistic.

In general, the discrete hazards approach performed well, though perhaps not sufficiently better than some of the proposed parametric models (such as lognormal or piecewise exponential) to justify the need for six parameters in estimating the survival curve. Furthermore, assuming a constant hazard over certain age intervals is not necessarily an assumption we wish to make, as it is unrealistic even at a fine scale of age groups. Additional comments on the results of the application can be found in Appendix D.

5. Discussion

In this paper, we propose two novel approaches to estimating full survival curves for child mortality that are well suited to applications in LMICs: one parametric and one nonparametric, survey-weighted survival model. We detail existing methods that can be used to estimate full survival curves and explain how they fall short in this specific context.

Our application suggests that there are potentially very large differences in model fit between parametric distributions, with the Weibull distributions and Gompertz distributions generally providing the worst fit compared to the Turnbull estimator, in terms of capturing the survival curve under the age of 5. This suggests that in scenarios where data may be sparse, the proposed approach is sensitive to the choice of parametric hazard, and as such, parametric assumptions should be assessed via some form of model validation. In general, the lognormal model seems to fit the countries in our application reasonably well. We note that two of the three-parameter models we compared – the piecewise exponential and ETSP model – also adequately captured the survival curve provided by the Turnbull estimator, though the piecewise exponential model has the undesirable property of assuming constant hazards within prespecified age groups, and the ETSP model is computationally challenging to fit. We conclude that for our application, the lognormal model outperforms other parametric models in terms of the ability to capture the point estimate provided by the Turnbull estimator while only requiring two parameters to define the survival curve. We note that this choice takes into consideration the trade-off between parsimony and flexibility; with little data, stronger parametric assumptions, which may involve fewer parameters, may be needed to produce realistic estimates. This consideration is especially important in small area estimation, where sample sizes are even more limited.

The benefits of a parametric approach to under-5 mortality estimation, and in particular to estimating the full survival curve for children under the age of 5, are many. As laid out in Schöley (2019), correctly specified parametric assumptions about the shape of mortality may greatly assist estimation of the survival curve under the age of 5 in situations with little data. This becomes particularly relevant in a small area setting, where often little data is available at small administrative regions (Wakefield, Okonek, and Pedersen 2020). The methods proposed in this paper may be used for small area estimation of child mortality when a full survival curve is desired, since they contain fewer parameters than in the usual discrete hazards approach (Mercer et al. 2015).

Further benefits of a continuous parametric approach involve interpretability and parsimony. The Heligman-Pollard model (Heligman and Pollard 1980), a well-known parametric demographic model for mortality estimation, provides informative interpretations of the parameters involved in the model, and the same is true of the models we propose. Of course, we rely on the assumption that the parametric distribution used is

correctly specified, which may not be the case. However, we have shown that our parametric approach may perform as good or better than the discrete hazards approach, and further exploration of parametric hazards should be investigated to see if model fit can be improved. We emphasize that our approach is generalizable to other parametric families, not only the ones considered in this paper, and that there may not be one single parametric distribution that best fits all countries. Especially in scenarios with little data, reasonable parametric assumptions are useful. Hence it is important to observe and test these parametric models in settings with more data, such as the national setting we use in our application. A meaningful question is: 'Is the fit of a continuous parametric model at least as good as the six-parameter discrete hazards model currently used by the UN IGME and DHS?' When compared to the Turnbull estimator, the lognormal model does outperform the six-parameter discrete hazards model in terms of our model performance metric (see Table 2).

An additional benefit of our approach is that we can address potential age heaping without the need to remove any data from our modeling procedure. Similarly to Romero Prieto, Verhulst, and Guillot (2021), we did not find large differences in resulting model estimates comparing models where we adjusted for age heaping to those where we did not (see Appendix E). Despite not seeing a large impact on estimates, we still find it valuable that our approach can incorporate this information, particularly since our application was not exhaustive and we cannot guarantee this would be the case across all DHS surveys or scenarios where age heaping may occur more generally.

Limitations of our nonparametric proposed approach include that the Turnbull estimator lacks a well-justified variance estimate. As previously noted, a variance estimate is not readily available due to the non-Gaussian, cubed-root asymptotics, and a bootstrap estimate of the variance is not applicable for similar reasons. More work needs to be done before comparisons between the nonparametric and parametric approaches (and model validation procedures) can be made with some degree of calibration. Alternative approaches to model validation such as evaluation via proper scoring rules could be considered, though existing approaches are not yet generalized to allow for both arbitrarily censored and truncated observations in addition to a complex survey design (Lumley and Scott 2015; Yanagisawa 2023).

In conclusion, we have provided a flexible framework for obtaining a complete continuous survival curve for children under the age of 5 using parametric models. Our method enables estimation using interval censored, left-truncated observations, as is required for period estimates of mortality from DHS data. Our approach is flexible in its ability to adapt to various shapes of mortality curves using different parametric hazard forms, and also allows for straightforward incorporation of covariates if this were desired. Furthermore, aspects of survey design, which are particularly relevant in LMICs, are directly acknowledged in our modeling framework to provide design-consistent estimates of mortality with finite population variances. Possible extensions of this work – in

addition to further exploration of parametric families – include subnational or small area estimation with the use of Bayesian smoothing models to incorporate spatial information into the estimation framework (Rao and Molina 2015; Wakefield, Okonek, and Pedersen 2020).

6. Acknowledgments

We are grateful to the editor and two anonymous reviewers for their helpful comments. This work was supported by NIH grant R01 HD112421-01. Views expressed in this article are those of the authors and do not necessarily reflect those of NIH.

References

Alexander, M., Zagheni, E., and Barbieri, M. (2017). A flexible Bayesian model for estimating subnational mortality. *Demography* 54(6): 2025–2041. doi:10.1007/s13524-017-0618-7.

Alkema, L. and New, J.R. (2014). Global estimation of child mortality using a Bayesian B-spline bias-reduction model. *The Annals of Applied Statistics* 8(4): 2122–2149. doi:10.1214/14-AOAS768.

Allison, P.D. (2014). *Event history and survival analysis: Regression for longitudinal event data*. Thousand Oaks: SAGE publications. doi:10.4135/9781452270029.

Barbieri, M., Wilmoth, J.R., Shkolnikov, V.M., Glei, D., Jasilionis, D., Jdanov, D., Boe, C., Riffe, T., Grigoriev, P., and Winant, C. (2015). Data resource profile: The Human Mortality Database (HMD). *International Journal of Epidemiology* 44(5): 1549–1556. doi:10.1093/ije/dyv105.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51(3): 279–292. doi:10.2307/1402588.

Binder, D.A. (1992). Fitting Cox's proportional hazards models from survey data. *Biometrika* 79(1): 139–147. doi:10.1093/biomet/79.1.139.

Breslow, N.E. and Wellner, J.A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics* 34(1): 86–102. doi:10.1111/j.1467-9469.2006.00523.x.

Breslow, N.E. and Wellner, J.A. (2008). AZ-theorem with estimated nuisance parameters and correction note for 'Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression'. *Scandinavian Journal of Statistics* 35(1): 186–192. doi:10.1111/j.1467-9469.2007.00574.x.

Carstensen, B. (2007). Age–period–cohort models for the Lexis diagram. *Statistics in Medicine* 26(15): 3018–3045. doi:10.1002/sim.2764.

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.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B* 39(1): 1–22.

Eilerts, H., Prieto, J.R., Eaton, J.W., and Reniers, G. (2021). Age patterns of under-5 mortality in sub-Saharan Africa during 1990–2018: A comparison of estimates from

demographic surveillance with full birth histories and the historic record. *Demographic Research* 44(18): 415–442. doi:10.4054/DemRes.2021.44.18.

Groeneboom, P. and Wellner, J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation*. Basel: Springer Science & Business Media. doi:10.1007/978-3-0348-8621-5.

Guillot, M., Romero Prieto, J., Verhulst, A., and Gerland, P. (2022). Modeling age patterns of under-5 mortality: Results from a log-quadratic model applied to high-quality vital registration data. *Demography* 59(1): 321–347. doi:10.1215/00703370-9709538.

Heligman, L. and Pollard, J.H. (1980). The age pattern of mortality. *Journal of the Institute of Actuaries* 107(1): 49–80. doi:10.1017/S0020268100040257.

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

Kaplan, E.L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association* 53(282): 457–481. doi:10.2307/2281868.

Lawn, J.E., Osrin, D., Adler, A., and 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.

Li, Z., Hsiao, Y., Godwin, J., Martin, B.D., Wakefield, J., and Clark, S.J. (2019). Changes in the spatial distribution of the under-five mortality rate: Small-area analysis of 122 DHS surveys in 262 subregions of 35 countries in Africa. *PloS One* 14(1): e0210645. doi:10.1371/journal.pone.0210645.

Lin, D. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* 87(1): 37–47. doi:10.1093/biomet/87.1.37.

Lumley, T. and Scott, A. (2015). AIC and BIC for modeling with complex survey data. *Journal of Survey Statistics and Methodology* 3(1): 1–18. doi:10.1093/jssam/smu021.

Mercer, L.D., Wakefield, J., Pantazis, A., Lutambi, A.M., Masanja, H., and Clark, S. (2015). Space-time smoothing of complex survey data: Small area estimation for child mortality. *The Annals of Applied Statistics* 9(4): 1889–1905. doi:10.1214/15-AOAS872.

Pedersen, J. and Liu, J. (2012). Child mortality estimation: Appropriate time periods for child mortality estimates from full birth histories. *PLoS Medicine* 9(8). doi:10.1371/journal.pmed.1001289.

Prentice, R.L. (1974). A log gamma model and its maximum likelihood estimation. *Biometrika* 61(3): 539–544. doi:10.1093/biomet/61.3.539.

Rao, J.N. and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association* 83(401): 231–241. doi:10.2307/2288945.

Rao, J.N. and Molina, I. (2015). *Small area estimation*. Hoboken: John Wiley & Sons.

Romero Prieto, J., Verhulst, A., and Guillot, M. (2021). Estimating the infant mortality rate from DHS birth histories in the presence of age heaping. *PLoS One* 16(11): e0259304. doi:10.1371/journal.pone.0259304.

Schöley, J. (2019). The age-trajectory of infant mortality in the United States: Parametric models and generative mechanisms. Abstract presented at Annual Meeting of the Population Association of America, Austin, TX, April 10–13, 2019.

Sun, J. (2001). Variance estimation of a survival function for interval-censored survival data. *Statistics in Medicine* 20(8): 1249–1257. doi:10.1002/sim.719.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society: Series B* 38(3): 290–295.

Verhulst, A., Prieto, J.R., Alam, N., Eilerts-Spinelli, H., Erchick, D.J., Gerland, P., Katz, J., Lankoande, B., Liu, L., Pison, G., Reniers, G., Subedi, S., Villavicencio, F., and Guillot, M. (2022). Divergent age patterns of under-5 mortality in south Asia and sub-Saharan Africa: A modelling study. *The Lancet Global Health* 10(11): e1566–e1574. doi:10.1016/S2214-109X(22)00337-0.

Wakefield, J., Okonek, T., and Pedersen, J. (2020). Small area estimation for disease prevalence mapping. *International Statistical Review* 88(2): 398–418. doi:10.1111/insr.12400.

Wang, Z., Gardiner, J., and Ramamoorthi, R. (1994). Identifiability in interval censorship models. *Statistics and Probability Letters* 21(3): 215–221. doi:10.1016/0167-7152(94)90117-1.

Wilmoth, J., Zureick, S., Canudas-Romo, V., Inoue, M., and 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.

Wu, Y., Li, Z.R., Mayala, B.K., Wang, H., Gao, P., Paige, J., Fuglstad, G.A., Moe, C., Godwin, J., Donohue, R.E., Croft, T.N., and Wakefield, J. (2021). Spatial modeling for subnational administrative level 2 small-area estimation. DHS Spatial Analysis Reports 21. Rockville, USA: ICF.

Yanagisawa, H. (2023). Proper scoring rules for survival analysis. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.). *Proceedings of the 40th International Conference on Machine Learning*. PMLR, Proceedings of Machine Learning Research, 39165–39182.

Appendices

A. Data

A.1 Survey Design

All DHS surveys used in our application follow a two-stage stratified cluster design and were designed to provide accurate estimates at the Administrative 1 (admin1) subnational level. Strata are defined by admin1 region and urban/rural status. Each sampling frame is established from a previous census. Primary sampling units (PSUs), or clusters, are selected across strata, and the second stage of sampling consists of households within PSUs. GPS coordinates are displaced by up to 2km for urban clusters and 5km for rural clusters, and are not displaced outside of their strata. Information related to the sampling design for the surveys used in our application is given in Table A-1.

Table A-1: Sampling information from DHS surveys

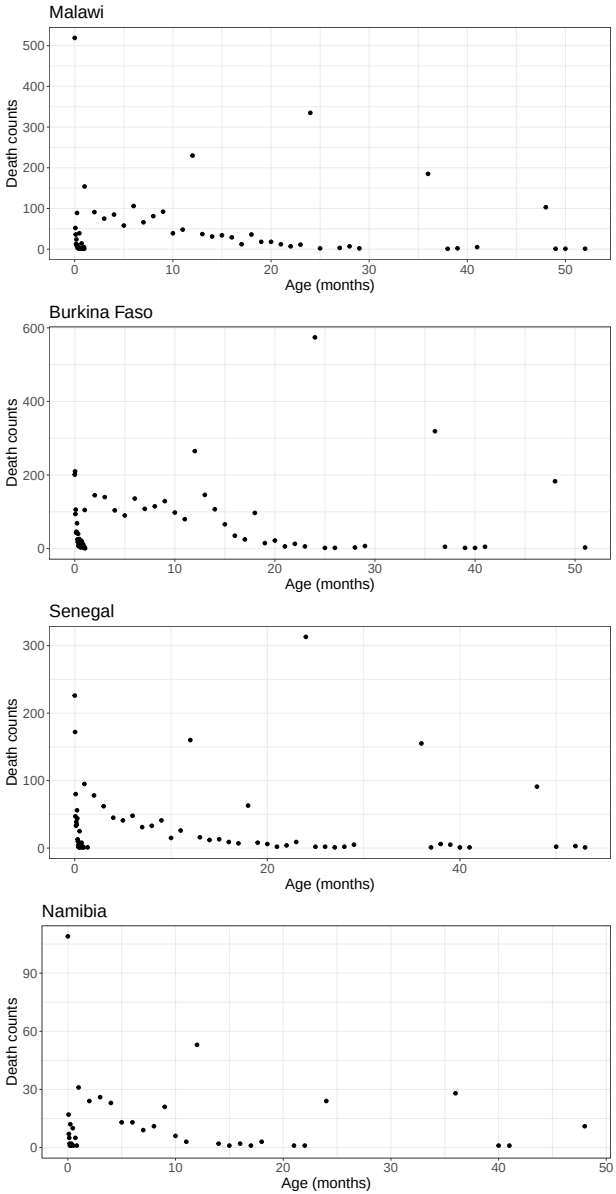
Country	Survey Year	Census	Admin1 Regions	PSUs (U/R)	Households (U/R)
Burkina Faso	2010	2006	13	574 (176/398)	14,924 (4576/10348)
Malawi	2016	2008	*28	850 (173/677)	27,531 (5190/22341)
Senegal	2010	2002	14	392 (147/245)	8,232 (3087/5145)
Namibia	2013	2011	13	554 (269/285)	11,080 (5380/5700)

Notes: Census year is the year of the census that the sampling frame for the survey is based upon. PSUs and Households listed are the number of PSUs and Households in the sample, not the sampling frame, and counts are additionally disaggregated by Urban/Rural (U/R). *At the time of survey, Malawi's 28 districts were considered Admin2 regions, with Northern, Central, and Southern regions being Admin1. Some shapefiles now consider the 28 districts to be Admin1, with a finer grid as 243 Admin2 subregions.

A.2 Age heaping in DHS surveys

For the four DHS surveys we consider in our application (Malawi 2015–2016, Burkina Faso 2010, Senegal 2010, Namibia 2013), we display the total death counts at each age recorded in the entire survey in Figure A-1. Note that we expect peaks at 24, 36, and 48 months because they capture a full year of deaths as opposed to only single months, but the peaks observed at 12 months reflect age heaping as they cover the same age span as the age groups surrounding it. The small number of counts observed at unexpected age months (25 months, for example) are the few exceptions to the typical interval censoring scheme used in DHS surveys.

Figure A-1: Total death counts at each age group within DHS surveys



B. Pseudo-likelihood

Consider a vector of superpopulation parameters $\boldsymbol{\theta}$ and observations $\mathbf{y}$ that can be written as the solution to the score equations $\mathcal{S}(\boldsymbol{\theta}, \mathbf{y}) = \mathbf{0}$. In a finite population setting, we are interested in the finite population parameters $\boldsymbol{\theta}'$, obtained by solving $\sum_{i=1}^{N} \mathcal{S}(\boldsymbol{\theta}', \mathbf{y}) = \mathbf{0}$, where $i = 1, \ldots, N$ denote all individuals in the finite population. Rather than observe all N individuals in the population, we instead take a probability sample of $j = 1, \ldots, U \leq N$ individuals, with weights w_j equal to the inverse of their inclusion probabilities in the sample. We obtain survey-weighted estimates of the finite population parameters $\boldsymbol{\theta}'$ via the finite population score equations

$$\sum_{j=1}^{U} w_j \times \mathcal{S}(\boldsymbol{\theta}', \mathbf{y}_j) = \mathbf{0}.$$

This framework, developed in Binder (1983), works for sample designs and populations that admit asymptotically normal estimators, with certain regularity conditions (see Binder (1983) for details, as well as a more general case of score equations). As an example, for linear regression we would set $\mathcal{S}(\boldsymbol{\theta}, \mathbf{y}) = -(y_i - \mathbf{x}_i^\top \boldsymbol{\theta})\mathbf{x}_i$. The variance of $\boldsymbol{\theta}'$ then has a sandwich form obtained via a Taylor expansion of the score equations at $\boldsymbol{\theta}' = \boldsymbol{\theta}$.

C. Methodological details

C.1 Turnbull estimator

Suppose we are interested in estimating a full mortality schedule (survival curve) from ages 0 to 5 at a national level in time periods $[2000, 2005)$ and $[2005, 2010)$. For simplicity, we consider age in full years rather than months in this example. Let $[t_0, t_1]$ denote the interval in which an observation is censored, with the convention $[t_0, \infty)$ for right censored observations. Our data is recorded as follows:

Table A-2: (Left) Example data for children alive between the ages of 0 and 5 years in the time periods $[2000, 2005)$ and $[2005, 2010)$. (Right) Example data, separated into (truncated) observations for each time period

Individual	Year born	t_0	t_1
1	2002	2	4
2	1998	5	∞
3	2007	1	1
⋮	⋮	⋮	⋮

Individual	Year born	B_i	A_i	Period
1	2002	$(-\infty, \infty)$	[2, 3]	[2000, 2005)
1	2002	[3, ∞)	[3, 4]	[2005, 2010)
2	1998	[2, ∞)	[5, ∞)	[2000, 2005)
3	2007	$(-\infty, \infty)$	[1, 1]	[2005, 2010)
⋮	⋮	⋮	⋮	⋮

Individual 1 is interval censored to the age range $[2, 4]$, individual 2 is right censored at age 5, individual 3 is observed to die at exactly 1 year, and so forth. To account for time period as a time-varying covariate in our model, we rearrange our data by splitting each observation into multiple observations, one for each time period in which they contribute to the risk set. The observations are left truncated at the beginning of each time period by $\max(0,$ age at which they enter the time period), where we denote this truncation in set notation as B_i. If $B_i = (-\infty, \infty)$, this indicates that no truncation has occurred, which in this context means the individual was born in the given time period and not before it. Individuals are censored according to the sets $A_i = [L_i, R_i]$, where if $L_i = R_i$ the observation is exactly observed (uncensored), and if $A_i = [L_i, \infty)$ the observation is right censored at L_i. The number of observations in our expanded dataset is $N \equiv \sum_{i=1}^{n} p_i$, where n is the number of individuals in our dataset and p_i is the number of time periods in which individual i contributes to the risk set.

In the above example, for $i = 1, \ldots, N$ observations, let A_i denote an individual's censoring set, and let B_i denote an individuals truncation set such that the likelihood contribution for an individual can be denoted $P(X_i \in A_i \mid X_i \in B_i)$, where X_i is the random variable corresponding to the death of child i. The data is thus in the form of pairs $(A_1, B_1), \ldots, (A_n, B_n)$.

The Turnbull estimator is the NPMLE of the cumulative distribution function $\hat{F}$ of F, and therefore, also produces the NPMLE of the survival curve $\hat{S} = 1 - \hat{F}$. We write the likelihood in a convenient way that allows us to optimize the proportions of the cumulative distribution function that lie within given intervals subject to the constraint that the proportions sum to 1 and are greater than 0. This allows us to view the likelihood maximization as a constrained optimization problem that can be solved using an expectation-maximization algorithm (Dempster, Laird, and Rubin 1977), which we now describe.

Assume that each A_i can be written as a finite union of disjoint, closed intervals, with a single point $A_i = X_i$ being written as the closed interval $[X_i, X_i]$. Then for each censoring set A_i we can write,

$$A_i = \bigcup_{j=1}^{k_i} [L_{ij}, R_{ij}],$$

for $i = 1, \ldots, n$. Then the likelihood for all observations can be written as

$$\text{Likelihood} = \prod_{i=1}^{n} \frac{\left[\sum_{j=1}^{k_i} F(R_{ij}) - F(L_{ij})\right]}{P(B_i)}. \tag{3}$$

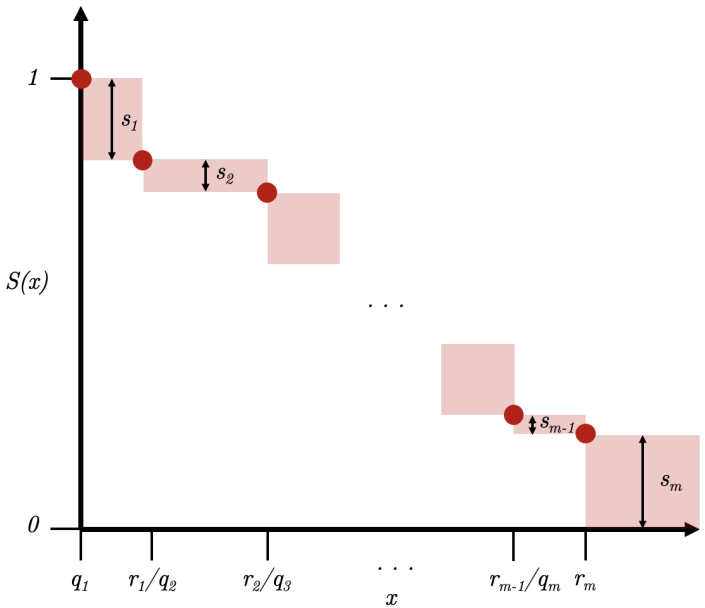
Let $[q_1, r_1], \ldots, [q_m, r_m]$ denote all unique intervals defined by $[L_{ij}, R_{ij}]$, and $s_j = F(r_j) - F(q_j)$. These s_j define the proportions of the cumulative distribution function that lie within an interval $[q_j, r_j]$.

We can rewrite the likelihood defined in Equation (3) as

$$\text{Likelihood} = \prod_{i=1}^{n} \left(\frac{\sum_{j=1}^{m} I\{[q_j, r_j] \in A_i\} s_j}{\sum_{j=1}^{m} I\{[q_j, r_j] \in B_i\} s_j} \right).$$

Maximizing the above subject to the constraints $s_j \geq 0$, $\sum_{j=1}^{m} s_j = 1$ then corresponds to maximizing the likelihood for arbitrarily censored and truncated observations, and provides us with a nonparametric estimate of the MLE. A visual description of the Turnbull estimator is provided in Figure A-2.

Figure A-2: A visual representation of the values s_j that make up the Turnbull estimator



Turnbull (1976) suggests the following procedure for obtaining the MLE:

1. Obtain initial values for $\mathbf{s}^0 = s_1^0, \ldots, s_m^0$ subject to $\sum_{j=1}^m s_j^0 = 1$, $s_j^0 \geq 0$.
2. Compute

$$\mu_{ij}(\mathbf{s}) = \frac{I\{[q_j, r_j] \in A_i\} s_j}{\sum_{k=1}^m, I\{[q_k, r_k] \in A_i\} s_k}$$
$$\nu_{ij}(\mathbf{s}) = \frac{(1 - I\{[q_j, r_j] \in B_i\}) s_j}{\sum_{k=1}^m I\{[q_k, r_k] \in B_i\} s_k}$$
$$\pi_j(\mathbf{s}) = \frac{\sum_{i=1}^n (\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s}))}{\sum_{i=1}^n \sum_{j=1}^m (\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s}))}.$$

3. Set $s_j^1 = \pi_j(\mathbf{s}^0)$.
4. Return to Step 1.

The procedure exits once some predetermined required tolerance is achieved.

Incorporating survey weights into the Turnbull estimator is straightforward, as the likelihood contribution for each individual is simply multiplied by their survey weight. This corresponds to altering Step 2 in the algorithm described, replacing $\pi_j(\mathbf{s})$ with $\tilde{\pi}_j(\mathbf{s})$, where $\tilde{\pi}_j(\mathbf{s})$ is

$$\tilde{\pi}_j(\mathbf{s}) = \frac{\sum_{i=1}^{n} w_i \left(\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})\right)}{\sum_{i=1}^{n} w_i \sum_{j=1}^{m} \left(\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})\right)},$$

and w_i is the survey weight for a given individual.

C.1.1 A note on the Turnbull estimator

When an individual is interval censored across the boundary of a time period, that individual cannot be split into separate observations that are left truncated at the beginning of a given time period. Intuitively, this would imply that a single individual could contribute more than one death to the risk set, which biases estimates of mortality upwards, and also overstates the amount of information present. When working on a 5-year period scale, few observations are interval censored across the boundary of a time period. For yearly periods (or shorter), we expect this to be a more prevalent occurrence. For 2000–2009 data from Malawi, roughly 0.3% of all individuals are interval censored across a time period boundary, which constitutes approximately 4% of all observed deaths throughout 2000–2009. Table 3 gives the exact breakdown of these percentages for the additional countries we consider in our application.

Table A-3: Percentages of individuals who are interval censored across a time period boundary out of all individuals at risk, and percentages of individuals who are interval censored across a time period boundary out of all observed deaths for 2000–2009

Country	Percent across boundary out of all individuals	Percent across boundary out of all deaths
Burkina Faso	0.7	6
Malawi	0.3	4
Namibia	0.1	3
Senegal	0.3	4

To account for an individual in our application who is interval censored across time period boundaries in the Turnbull estimator, we split the individual into separate left-

truncated observations at the beginning of the time period, where the last two observations for this individual will each contain an interval censored observation. We then down-weight these last two observations by the proportion of the length of the original individual's interval that is included in that time period. Note that this assumes that the age distribution of deaths in each of these two time periods is the same. Though we know this assumption will not hold (due to cohort effects of conflicts, for example), the resulting survey-weighted Turnbull estimator is still a useful comparator, as it is a nonparametric estimator that estimates rates more robustly than a parametric estimator.

C.2 Parametric estimator

Suppose we have children $i = 1, \ldots, n$. Let,

1. $p = 1, \ldots, P$: consecutive time periods, which may be single years or combinations of years (e.g., 1 or 5 year periods)
2. l_p: length of period p, measured in the same units as age of child
3. y_p: date at the start of time period p
4. b_i: child's date of birth
5. $a_{pi} = y_p - b_i$: the age the child would be at y_p
6. I_i: an indicator that child i is interval censored. If $I_i = 1$, child i is interval censored. If $I_i = 0$, child i is right censored or has an exact death time
7. E_i: an indicator that child i's death is exactly observed. If $E_i = 1$, then $I_i = 0$, and if $E_i = 0$, then I_i could be 0 or 1
8. t_i: child's age at right censoring or age at death
9. t_{0i}: child's age at beginning of interval censoring, if child is interval censored
10. t_{1i}: child's age at end of interval censoring, if child is interval censored
11. $\tilde{p}_i$: if $E_i = 1$, the period in which that child died
12. $U_{x_i}(p) = \{p : a_{pi} > -l_p, a_{pi} < x_i\}$. $U_{x_i}(p)$ is the set of periods for which child i is alive and at risk of dying, where x_i is one of t_i, t_{0i}, or t_{1i} where appropriate

Let $F_{\boldsymbol{\theta}}$ denote the cumulative distribution function for the specified parametric distribution, and $H_{\boldsymbol{\theta}}$ the corresponding cumulative hazard function, dependent on a set of unknown parameters $\boldsymbol{\theta}$. In the case of simple random sampling, the likelihood for all

individuals in our dataset across all time periods can be written as

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} L_i(\boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \left[1 - F_{\boldsymbol{\theta},i}(t_i)\right]^{1-I_i} \left[F_{\boldsymbol{\theta},i}(t_{1i}) - F_{\boldsymbol{\theta},i}(t_{0i})\right]^{I_i} \left[f_{\boldsymbol{\theta},i}(t_i)\right]^{E_i}, \\
&= \prod_{i=1}^{n} \underbrace{\left[\exp\left(-H_{\boldsymbol{\theta},i}(t_i)\right)\right]^{1-I_i}}_{\text{right censored}} \\
&\quad \times \underbrace{\left[\exp\left(-H_{\boldsymbol{\theta},i}(t_{0i})\right) - \exp\left(-H_{\boldsymbol{\theta},i}(t_{1i})\right)\right]^{I_i}}_{\text{interval censored}} \\
&\quad \times \underbrace{\left[\exp\left(-H_{\boldsymbol{\theta},i}(t_i)\right) h_{\boldsymbol{\theta},\tilde{p}_i}(t_i)\right]^{E_i}}_{\text{exact}},
\end{aligned}
$$

where

$$
H_{\boldsymbol{\theta},i}(x_i) = \sum_{U_{x_i}(p)} \int_{\max\{a_{p_i},0\}}^{\min\{x_i, a_{p_i}+l_p\}} h_{\boldsymbol{\theta},p}(u)du,
$$

and $h_{\boldsymbol{\theta},p}(u)$ is a period-specific hazard function for a specified parametric distribution.

C.3 Influence functions

For a generic parametric model (in a non survey context) with $j = 1, \ldots, J$ parameters, let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_J)$ denote the MLE. We can write $\hat{\theta}_j$ as asymptotically linear, meaning

$$
\hat{\theta}_j - \theta_j = \frac{1}{n} \sum_{i=1}^{n} \Delta_i + o_p(n^{-1/2}),
$$

with influence functions Δ_i given by

$$
\Delta_i = \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log(L_i(\hat{\boldsymbol{\theta}}))\right] \left[\mathbf{H}_{\log(L_i)}\right],
$$

where $\frac{\partial}{\partial \boldsymbol{\theta}} \log(L_i(\hat{\boldsymbol{\theta}}))$ is an $n \times J$ dimensional matrix of score functions for each individual $i = 1, \ldots n$ with respect to parameters $j = 1, \ldots J$, and $\mathbf{H}_{\log(L_i)}$ denotes the Hessian of

the log likelihood. Then for an influence function of a pseudo-MLE with weights w_i, we can write

$$\hat{\theta}_j - \theta_j = \frac{1}{n}\sum_{i=1}^{n} w_i \Delta_i + o_p(n^{-1/2}).$$

To estimate the finite population variance of $\hat{\theta}_j$, calculating the finite population variance of $\sum_{i=1}^{n} w_i \Delta_i$ corresponds to the Taylor-linearization method described in Binder (1983). The convenience here lies in that $\sum_{i=1}^{n} w_i \Delta_i$ is a survey total, and the influence functions are simple to obtain since we have a parametric model. In fact, only the gradient of the log likelihood evaluated at the pseudo-MLE and the Hessian obtained during optimization of the weighted log likelihood are needed to calculate the finite population variance.

Since our log likelihood is given by

$$\log(\mathrm{L}) = \sum_{i=1}^{n} \left[(1 - I_i)(-H_i(t_i)) + I_i \log(\exp(-H_i(t_{0i})) - \exp(-H_i(t_{1i})))\right]$$

we can then obtain the score for an individual i for each parameter θ_j as

$$\frac{\partial}{\partial \theta_j} \log(\mathrm{L})(\hat{\boldsymbol{\theta}}) =$$
$$(1 - I_i)(-\frac{\partial}{\partial \theta_j} H_i(t_i)) + I_i \left(\frac{-\exp(-H_i(t_{0i}))\frac{\partial}{\partial \theta_j} H_i(t_{0i}) + \exp(-H_i(t_{1i}))\frac{\partial}{\partial \theta_j} H_i(t_{1i})}{\exp(-H_i(t_{0i})) - \exp(-H_i(t_{1i}))} \right).$$

In practice, it is more computationally efficient to calculate the gradient analytically, though we may calculate the gradient numerically as well.

C.4 Log-quad model

The only parameter that is estimated in the modeling step when predicting the U5MR pattern for a new country is k, unless the average pattern of mortality observed across data in the HMD is desired, in which case k is set to 0. (Of note, Guillot et al. (2022) consider ${}_{60}\hat{q}_0$ to be an additional parameter for the log-quad model. Here we consider ${}_{60}\hat{q}_0$ to be data rather than a parameter, as ${}_{60}\hat{q}_0$ is input to the model as a fixed value rather than estimated during the modeling step.) The parameter k is estimated in one of two ways:

1. **Option 1**: If only a single value x is supplied for ${}_xq_0$ to the model as a fixed constant, k is estimated as

$$\hat{k} = \frac{e(x)}{v_x},$$

where $e(x)$ is the difference between the predicted and observed values of ${}_xq_0$ when the model is fit with $k = 0$.

2. **Option 2**: If more than one x is supplied for ${}_xq_0$ to the model as data, k is estimated as

$$\hat{k}^* = \frac{\sum_x w(x)e(x)v_x}{\sum_x w(x)v_x^2},$$

where $\hat{k}^*$ is the value of k that minimizes the root-mean-square error of predicted values of ${}_xq_0$ to observed values of ${}_xq_0$ across all values of x supplied, and $w(x)$ is the weight corresponding to the length of the previous age interval ending at age x (i.e., $w(1) = 7d$, $w(2) = 7d$, …, $w(22) = 1y$).

When all 22 possible values for x are supplied to the model, Guillot et al. (2022) propose an uncertainty band around the estimated survival curve that can be computed as

$$\hat{k}^* \pm 1.96 \times \sqrt{Var(\hat{k}^*)},$$
$$Var(\hat{k}^*) = \frac{22}{21}\left(\frac{\sum_x w(x)e(x)^2}{\sum_x w(x)v_x^2} - (\hat{k}^*)^2\right).$$

They propose a separate uncertainty band when only one value for x is supplied to the model, but we exclude it in our summary as that scenario is of little importance in our applications. In the derivation of this variance estimator, Guillot et al. (2022) assume that the errors $e(x)$ are independent across values x, that the weighted errors $w(x)e(x)$ are homoskedastic, and that $\hat{k}^*$ is approximately normally distributed.

Additionally, they note that almost all data used to estimate the age-specific coefficients $\{a_x, b_x, c_x, v_x\}$ in the U5MD is estimated with values of k that fall in $(-1, 1)$. Due to this observed range of values, they state that values of k that are estimated outside the range $(-1.1270, 1.5047)$ (the exact range of all observed values) have no "empirical basis" and may in fact produce estimates of ${}_xq_0$ that progress nonmonotically for children under the age of 5, whereas actual survival curves must be monotonically nonincreasing.

Therefore, they suggest a rule of thumb that the estimates should only be used when k is estimated in the range $(-1.1, 1.5)$.

C.5 Discrete hazards approach

We make two further notes on using the discrete hazards model in conjunction with DHS surveys. First, in DHS surveys deaths are recorded at exact days between ages 0 and 1 month, monthly until 24 months, and yearly onwards. The discrete hazards model does not take advantage of the fine-scale daily data available prior to 1 month, and instead groups those deaths together to form a neonatal age group. If NMR is the smallest demographic rate we wish to estimate, this grouping is not inherently an issue. However, daily recorded deaths may be informative of the overall pattern of mortality before the age of 5, so if we instead want to estimate an accurate survival curve over the entire age range from 0 to 5, grouping all deaths within the first month of life together will not capture the expected sharp decline in survival in the first week of life, or even the first two weeks of life.

A second benefit of aggregating our data across age groups is that, especially at small levels of spatial aggregation, we may have very little data available on the hazard in some age groups. Consequently, if data is sparse we may prefer to use fewer age groups in the discrete hazards model, as if no deaths were present in a certain age-period group (which is common for small regions in small area estimation, or fine-scale time periods), their estimated hazard will be exactly zero. Hazards of exactly zero are undesirable, as they are both implausible and it is difficult to get inference around such an estimate.

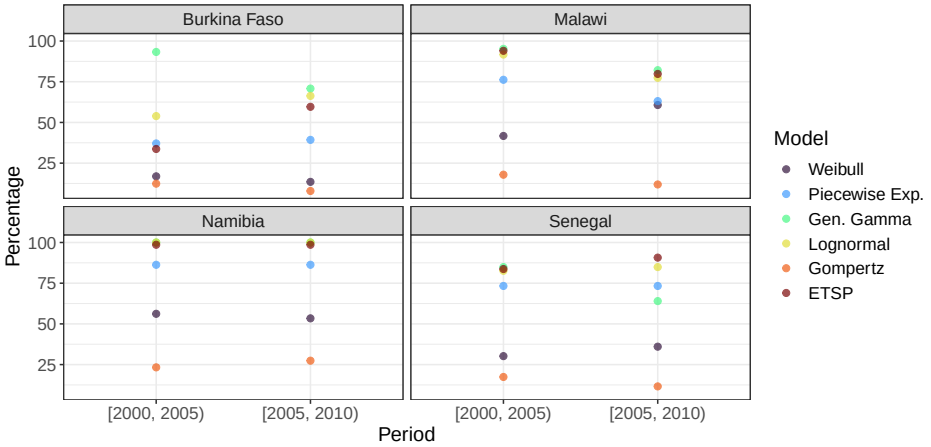
D. Additional results

One thing to note from the results of our application is that both the ETSP and generalized Gamma models performed reasonably well, but again perhaps not significantly better than some of the two-parameter models. This can be seen for Namibia in Table 2 in particular, where both differences between the models' respective parametric estimates and the Turnbull estimates capture zero for 99% (ETSP) and 100% (generalized Gamma) of ages where the Turnbull estimate is defined prior to 60 months. However, the lognormal model performs just as well in Namibia according to this metric. Furthermore, fitting both models (generalized Gamma and ETSP) results in computational complexities that may make them less desirable than other parametric options. The ETSP model, for example, does not have a closed form cumulative hazard, and therefore requires numerical integration at every step of the likelihood optimization. Therefore, this model takes more time to fit and is potentially less numerically stable than others with closed form cumulative

hazards. The generalized Gamma distribution, on the other hand, occasionally produces unreasonably wide confidence bands. In some countries and time periods, Burkina Faso $[2000, 2005)$, for example, the shape parameter Q is estimated with a large variance relative to the other parameters in the model. This results in the confidence bands produced being highly asymmetric, and so wide as to be unusable in practice. As such, we believe that in certain cases there may not be enough data under the age of 5 to reliably estimate all three parameters that define the generalized Gamma survival curve.

A visual representation of model validation results is presented in Figure A-3.

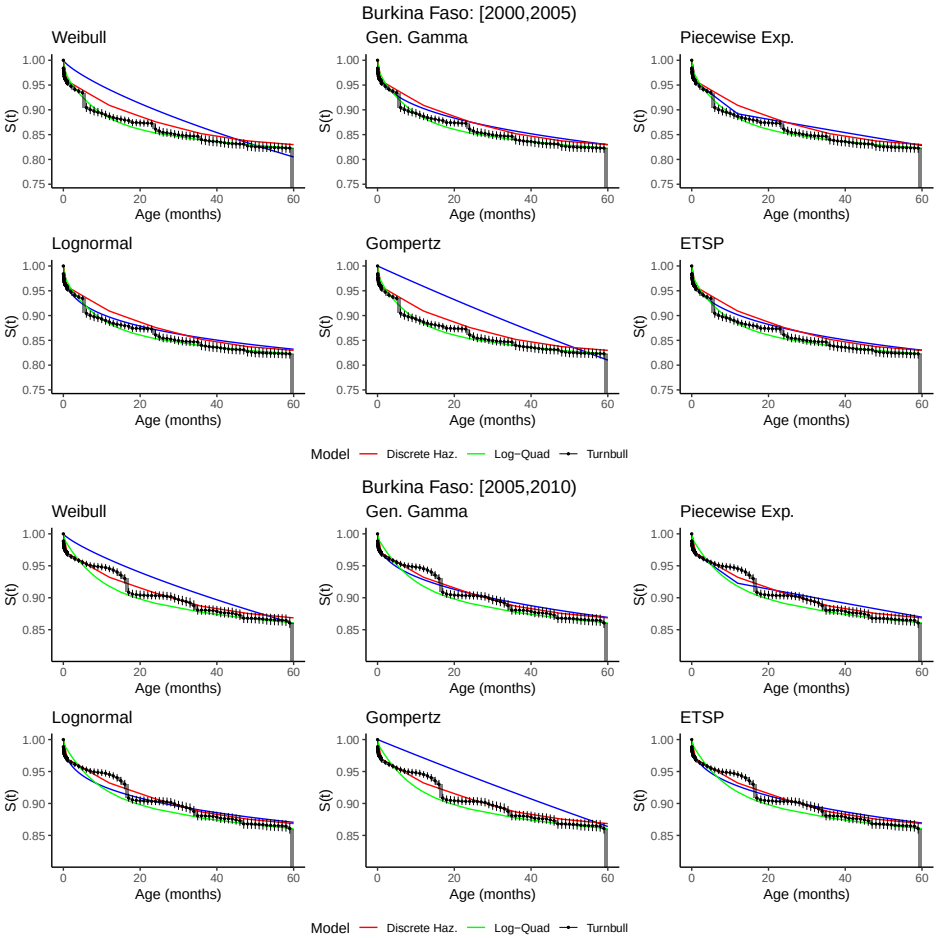
Figure A-3: Model validation results



Notes: Percentage of samples (out of 500) from $\hat{\theta} - \tilde{\theta}$ that contain 0 for all parametric models, countries, and periods.

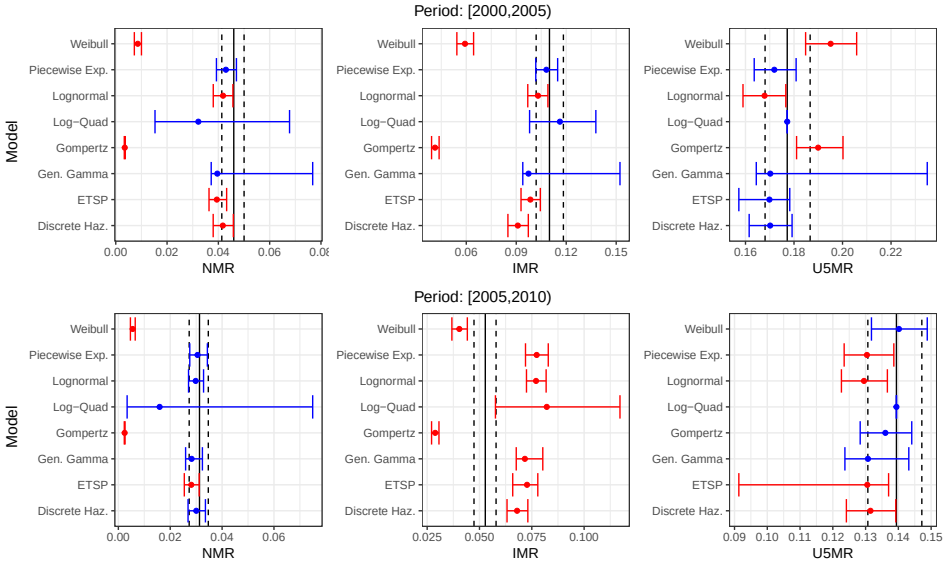
In the following section we display additional results from Weibull, generalized Gamma, piecewise exponential, lognormal, Gompertz, and exponentially truncated shifted power (ETSP) models for Burkina Faso, Malawi, Senegal, and Namibia.

Figure A-4: Estimated survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months



Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-5: Estimates of NMR, IMR, and U5MR for Burkina Faso in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)



Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-6: Empirical distributions of differences in survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$

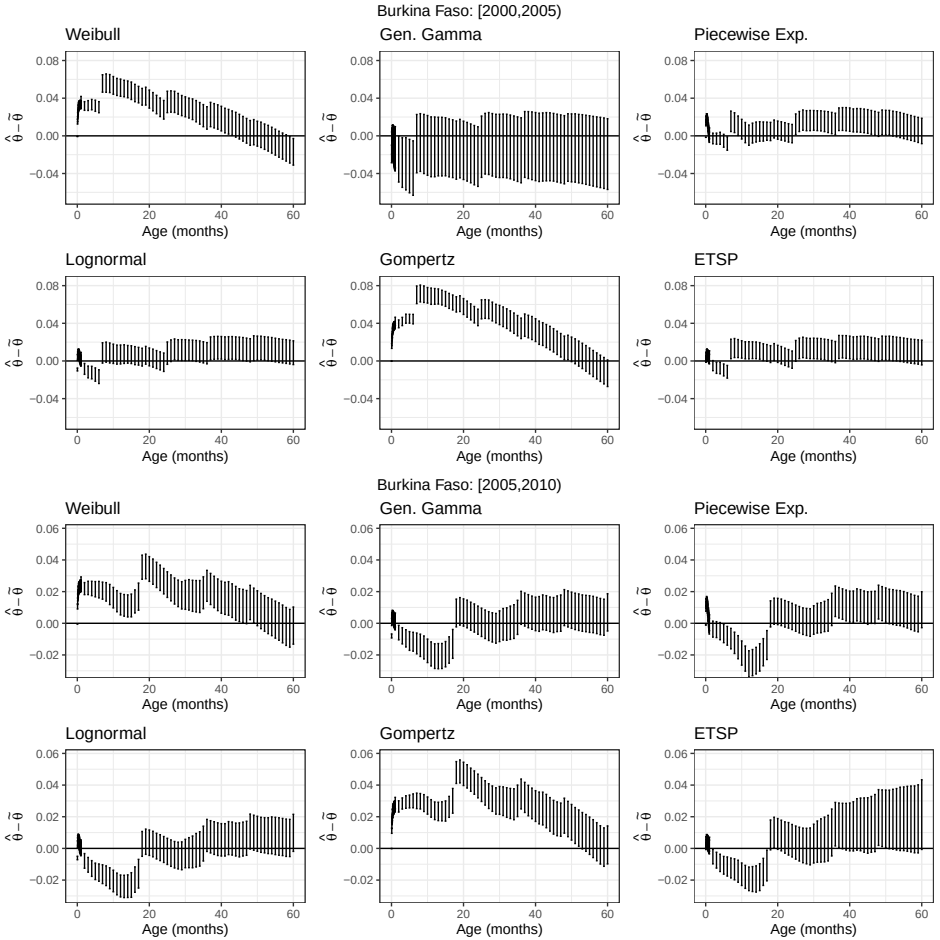
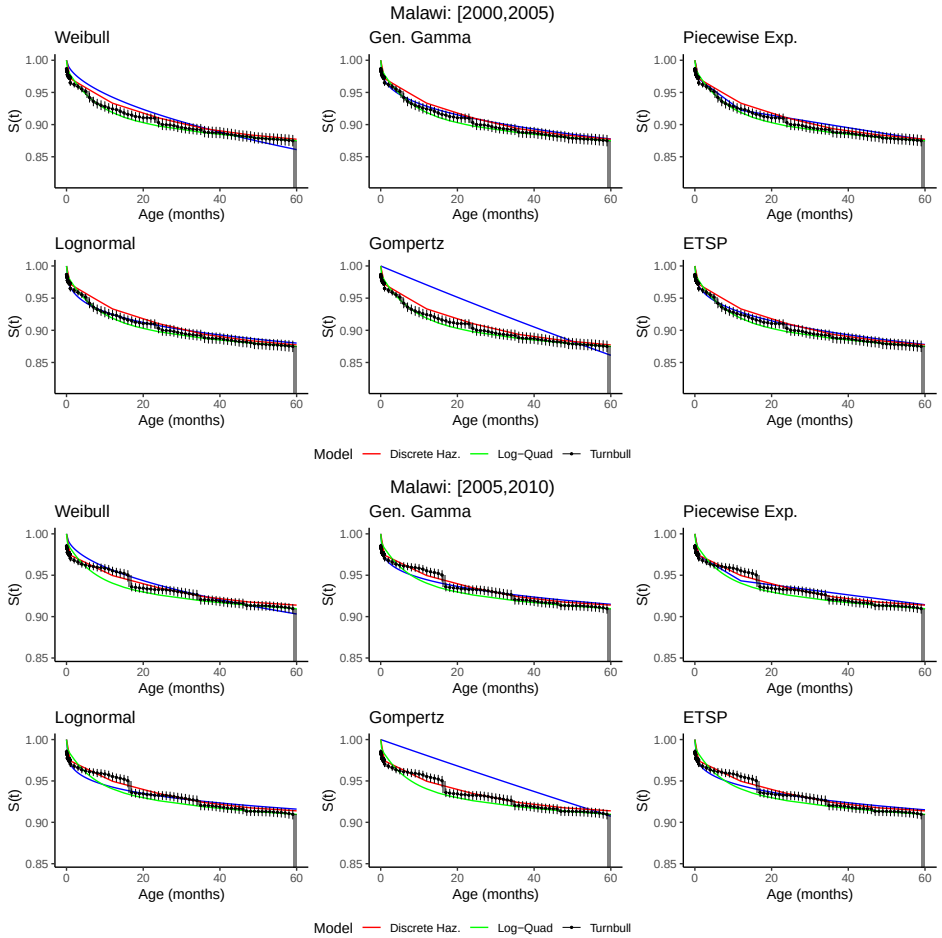
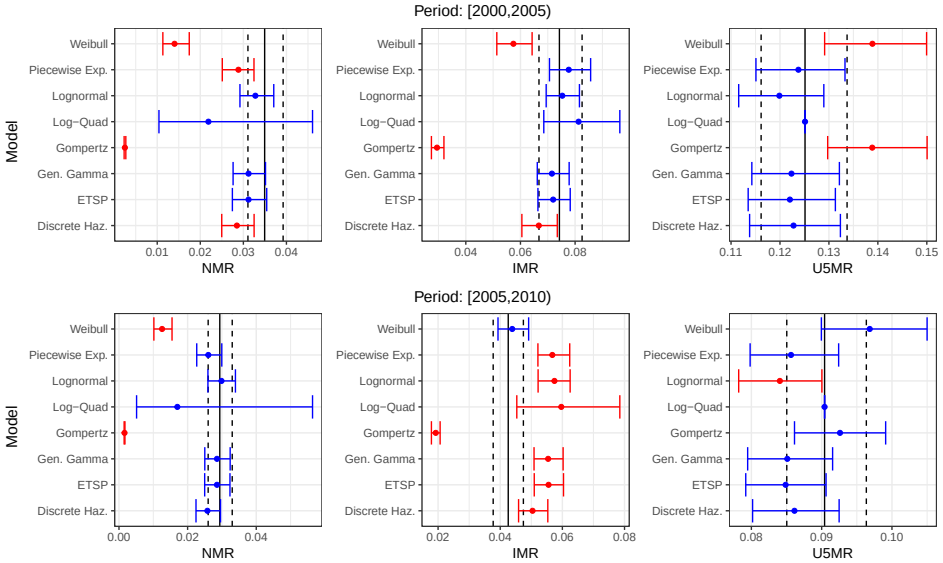


Figure A-7: Estimated survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months



Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-8: Estimates of NMR, IMR, and U5MR for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)



Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-9: Empirical distributions of differences in survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$

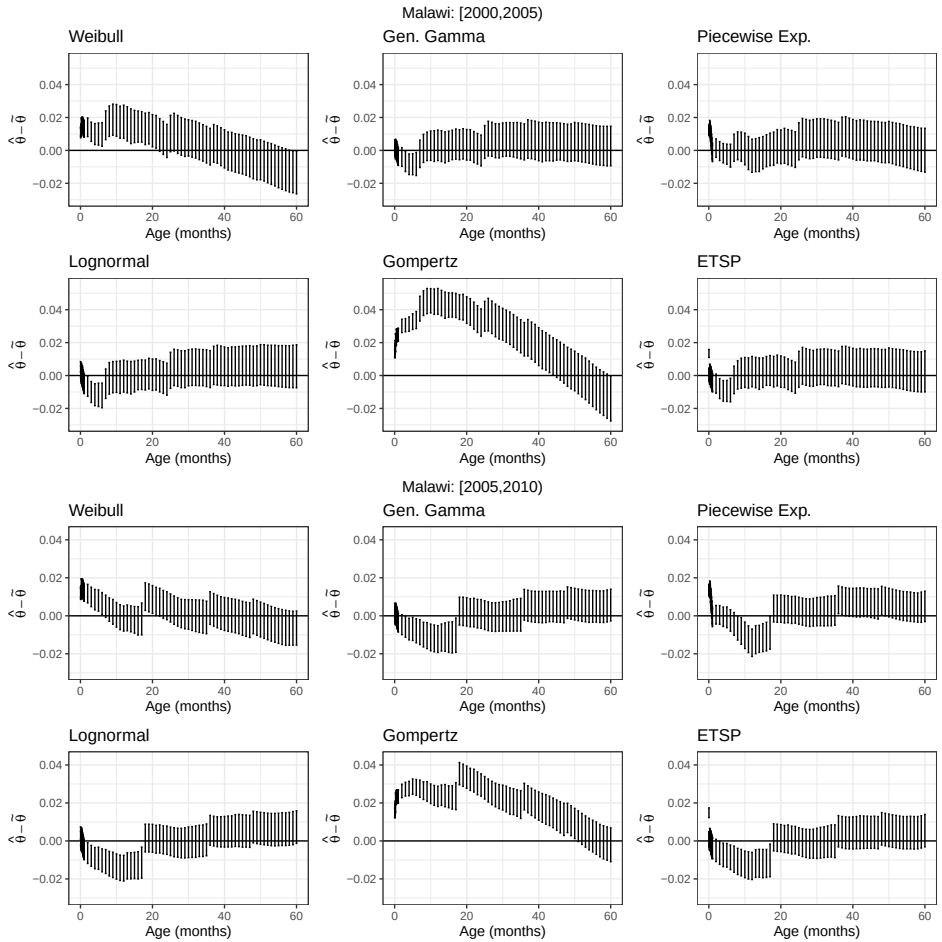
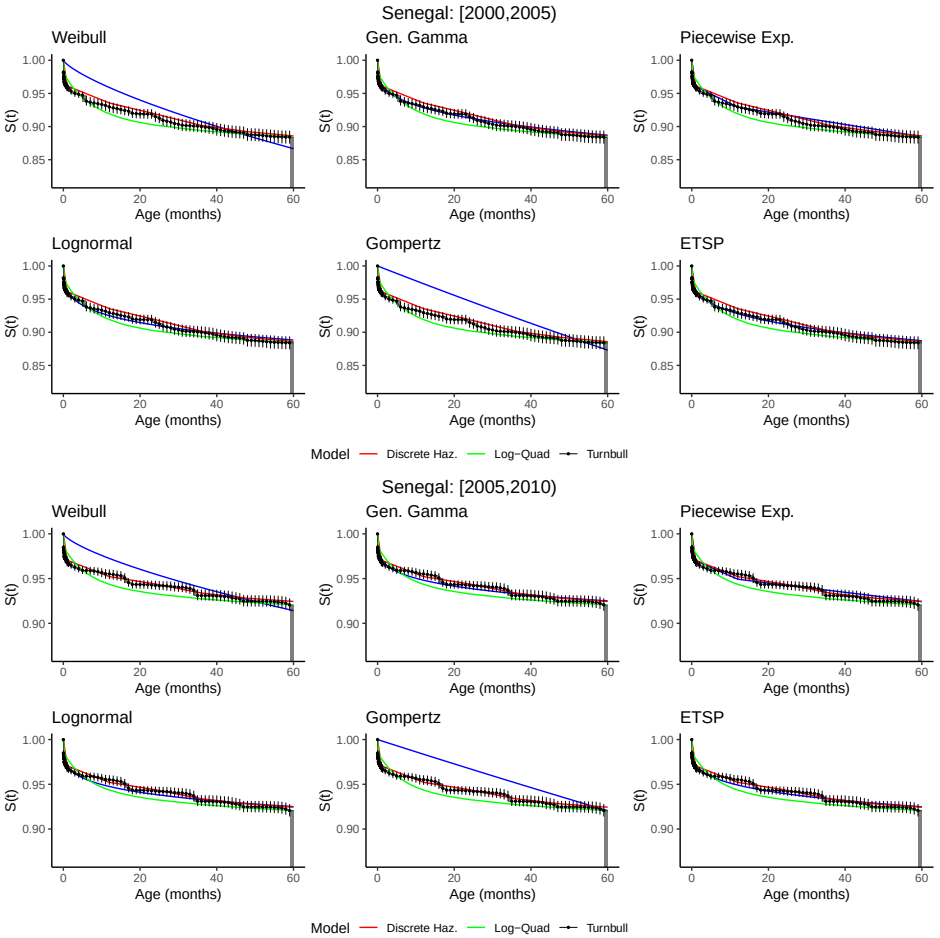
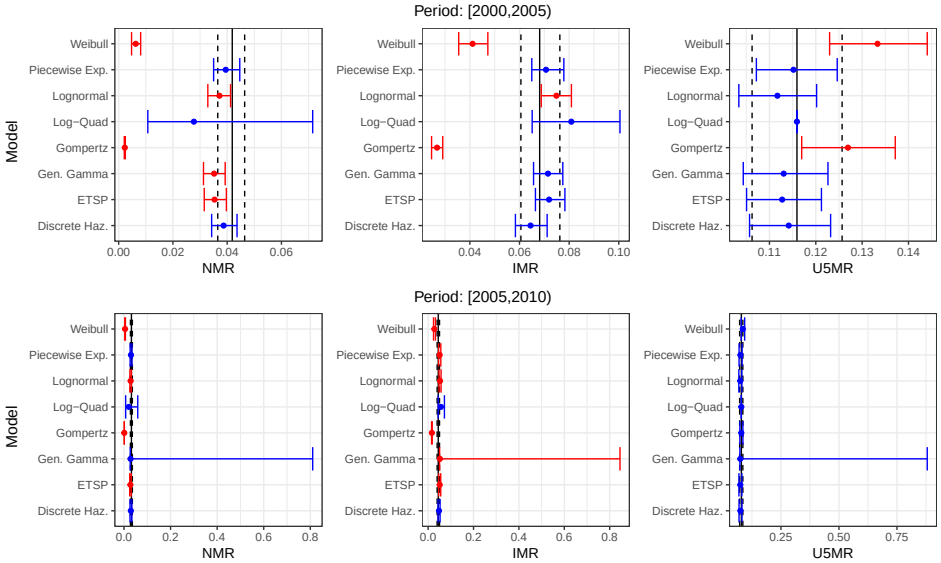


Figure A-10: Estimated survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months



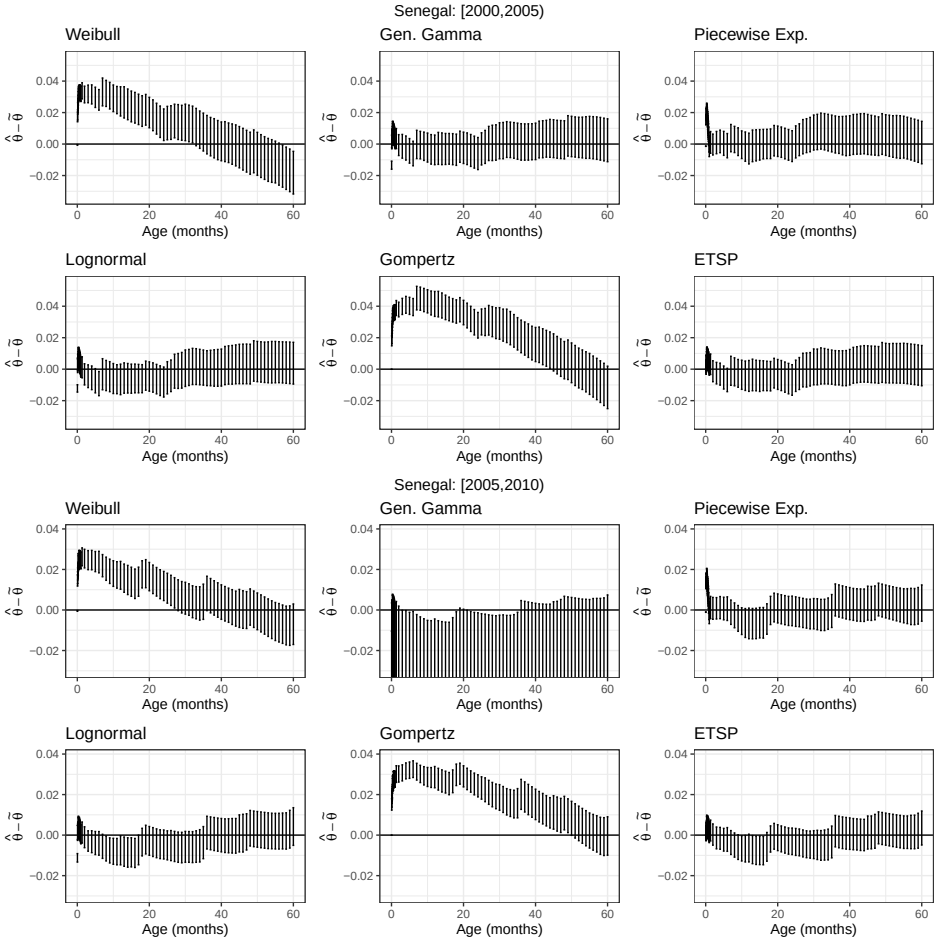
Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-11: Estimates of NMR, IMR, and U5MR for Senegal in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)



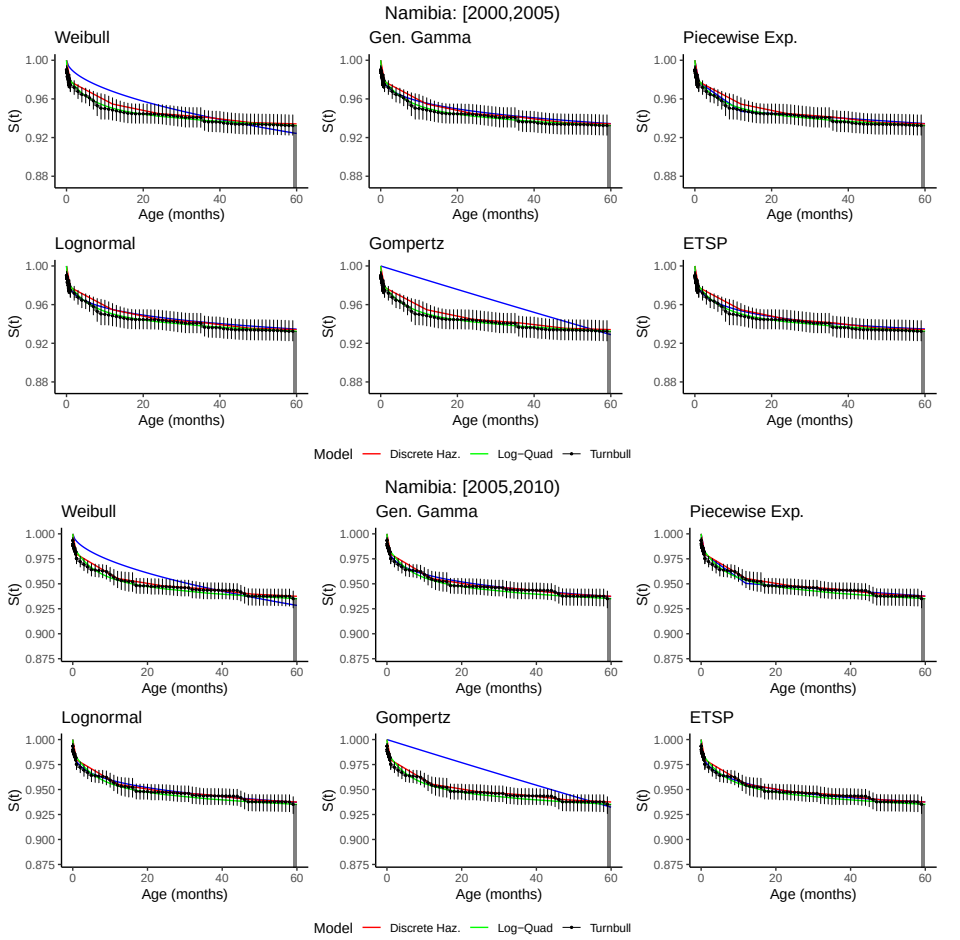
Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-12: Empirical distributions of differences in survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$



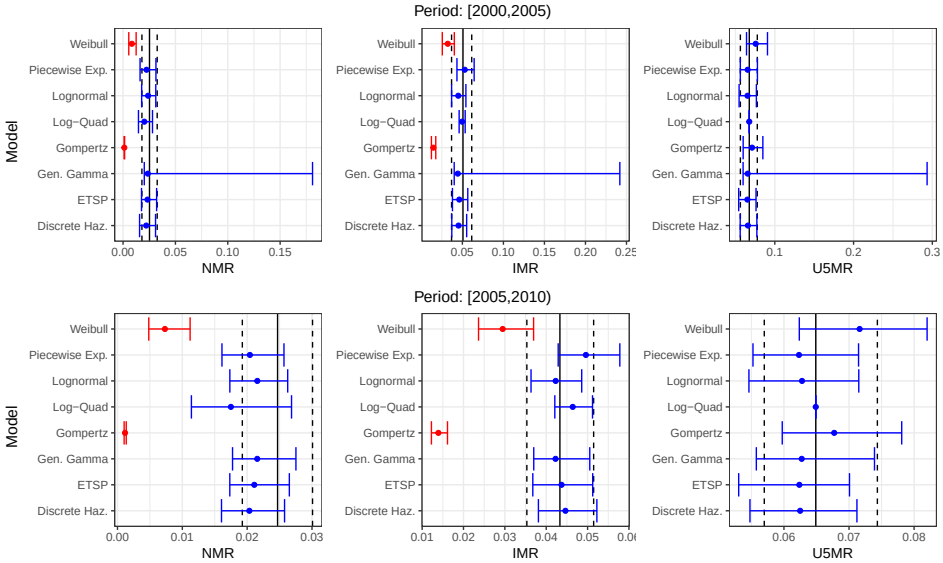
Notes: Note that for $[2005, 2010)$ the differences have been cut off at -0.03 for clarity, though the differences extend much further negative for the generalized gamma model.

Figure A-13: Estimated survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months



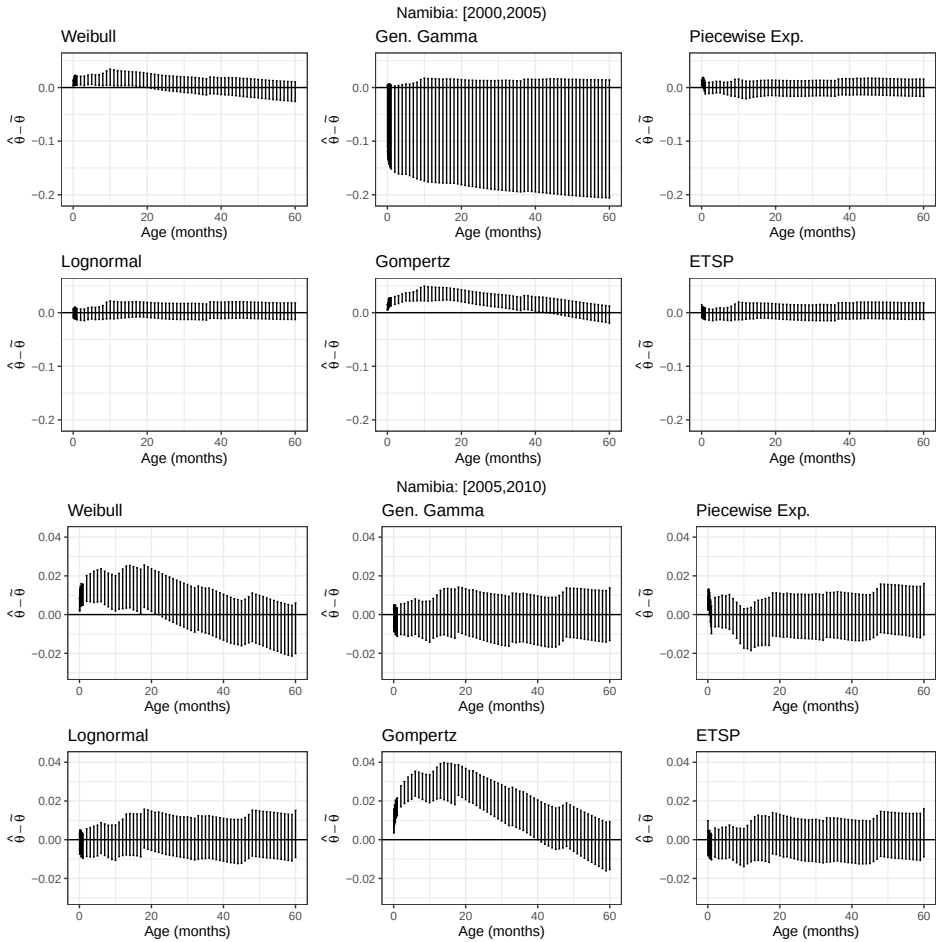
Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-14: Estimates of NMR, IMR, and U5MR for Namibia in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)



Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

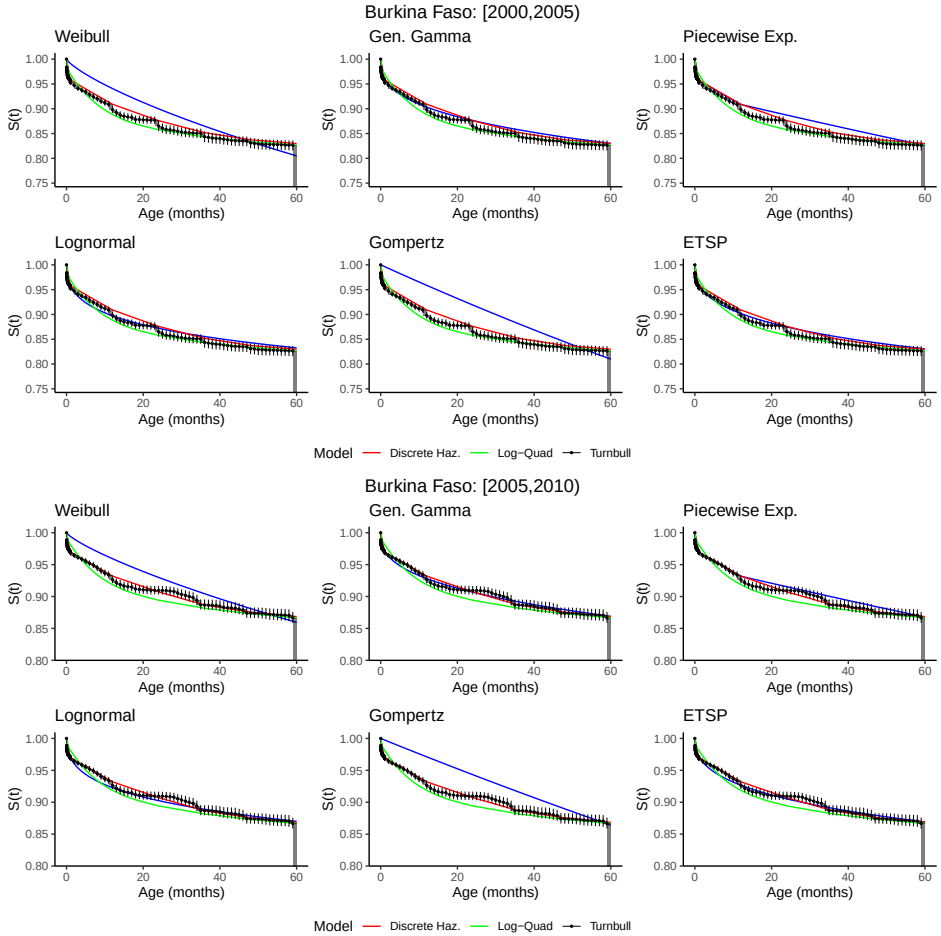
Figure A-15: Empirical distributions of differences in survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$



E. Comparison of models unadjusted for age heaping

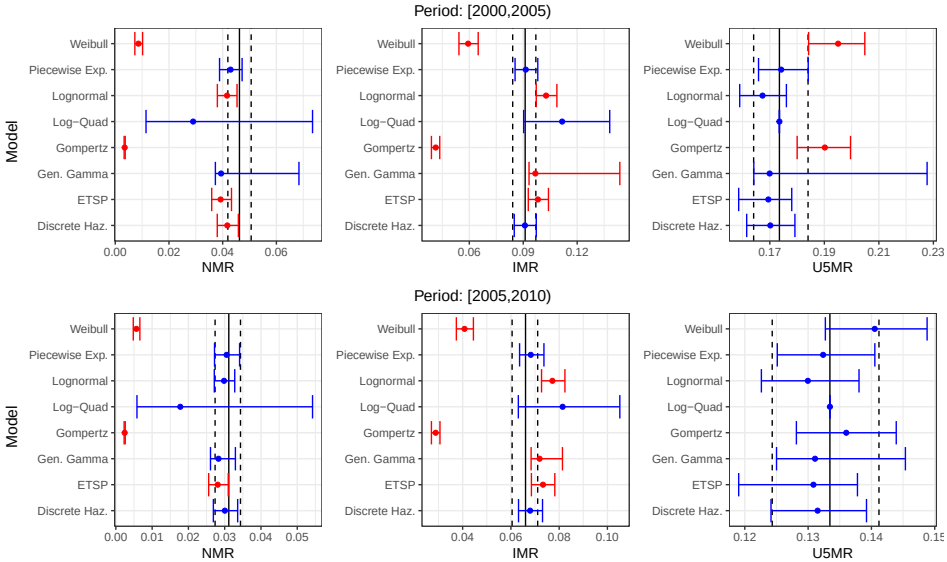
We repeat our application, fitting the same models without addressing age heaping at 12 months (by interval censoring observations recorded as having died between 6 and 18 months for that entire 12 month period).

Figure A-16: Estimated survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months



Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-17: Estimates of NMR, IMR, and U5MR for Burkina Faso in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months



Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-18: Empirical distributions of differences in survival curves for Burkina Faso in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$

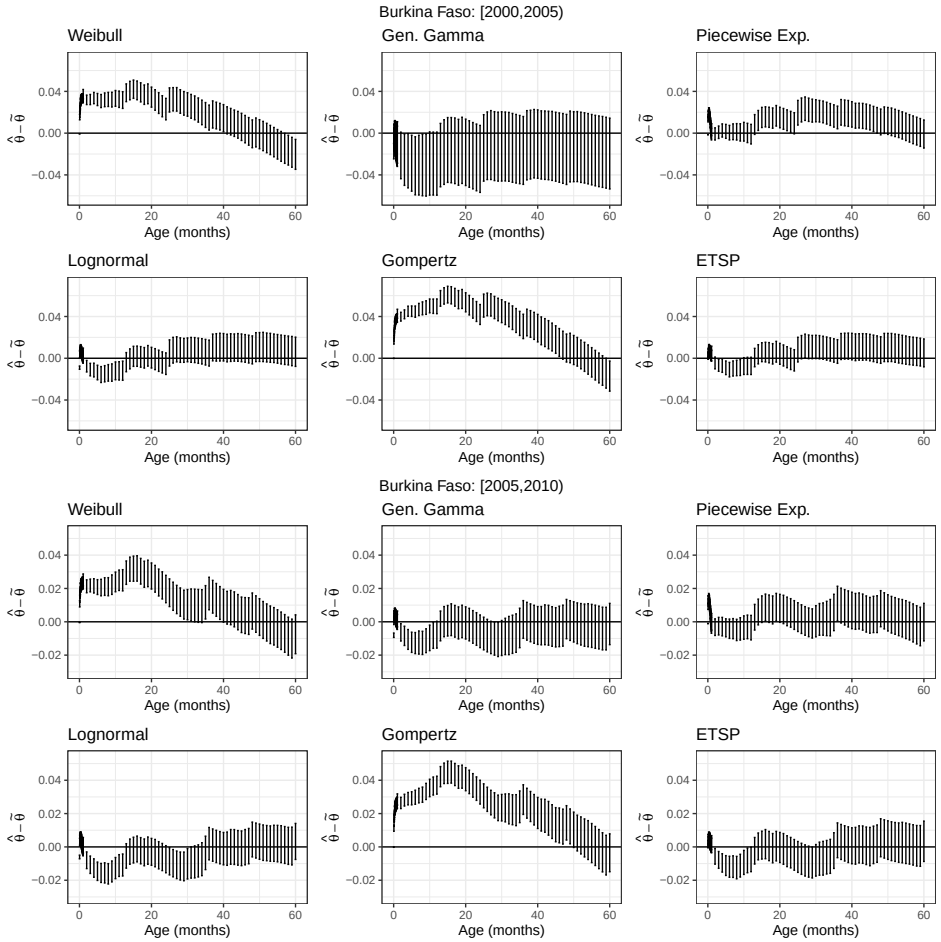
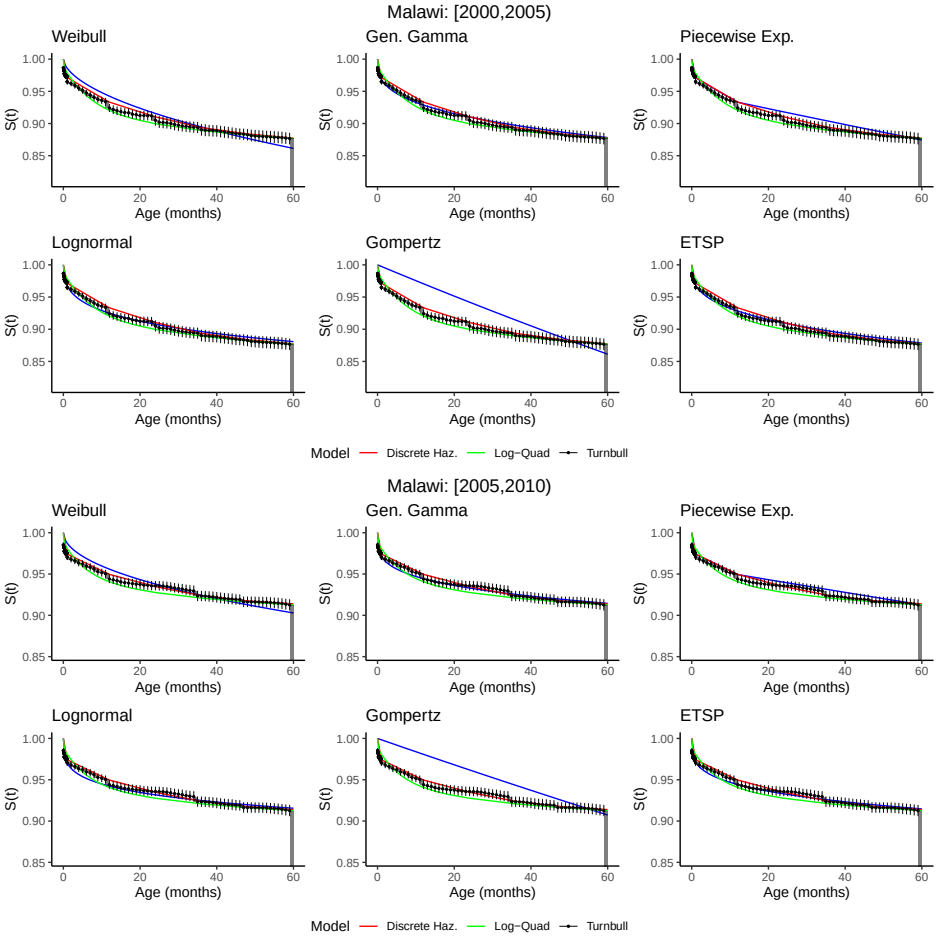
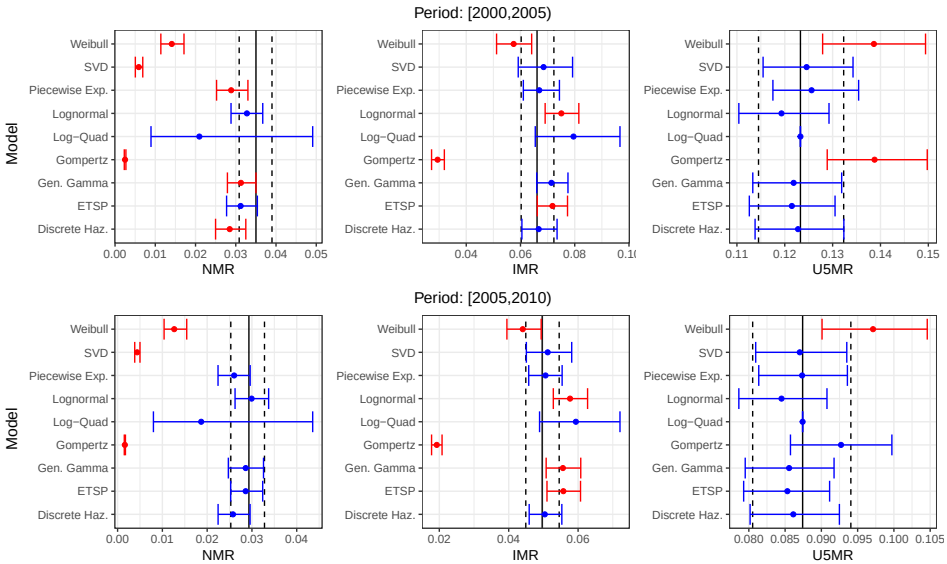


Figure A-19: Estimated survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months



Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-20: Estimates of NMR, IMR, and U5MR for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months



Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-21: Empirical distributions of differences in survival curves for Malawi in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$

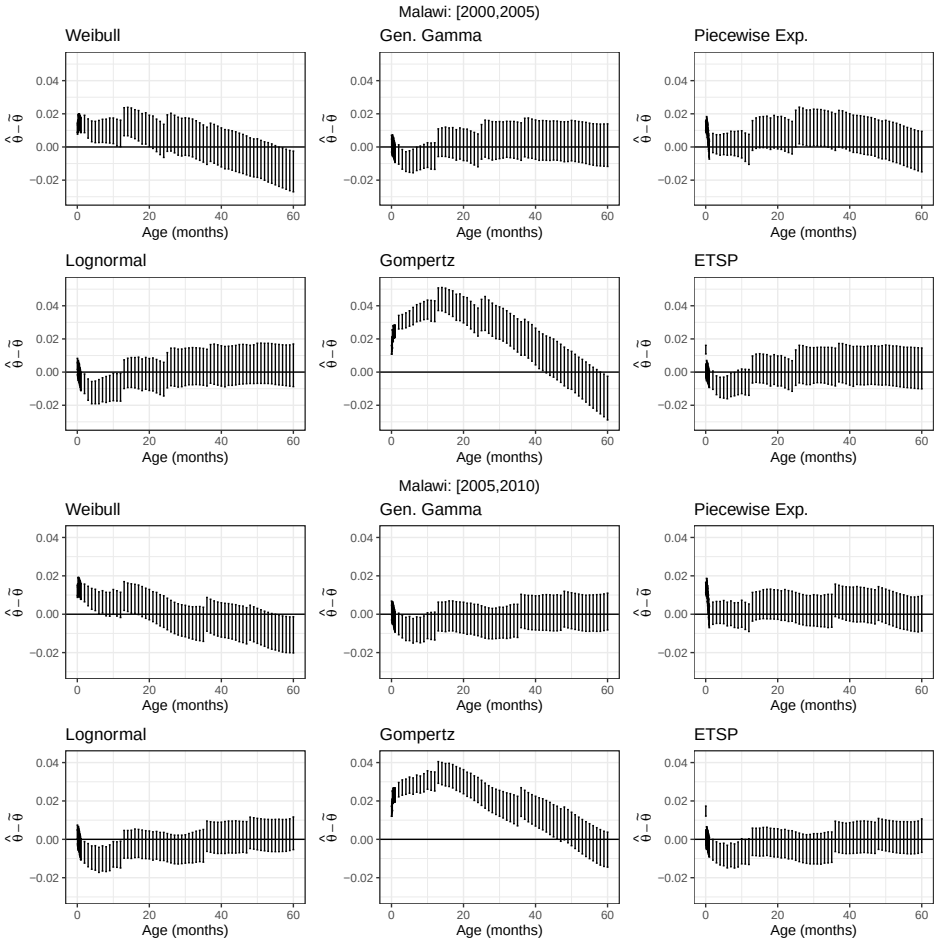
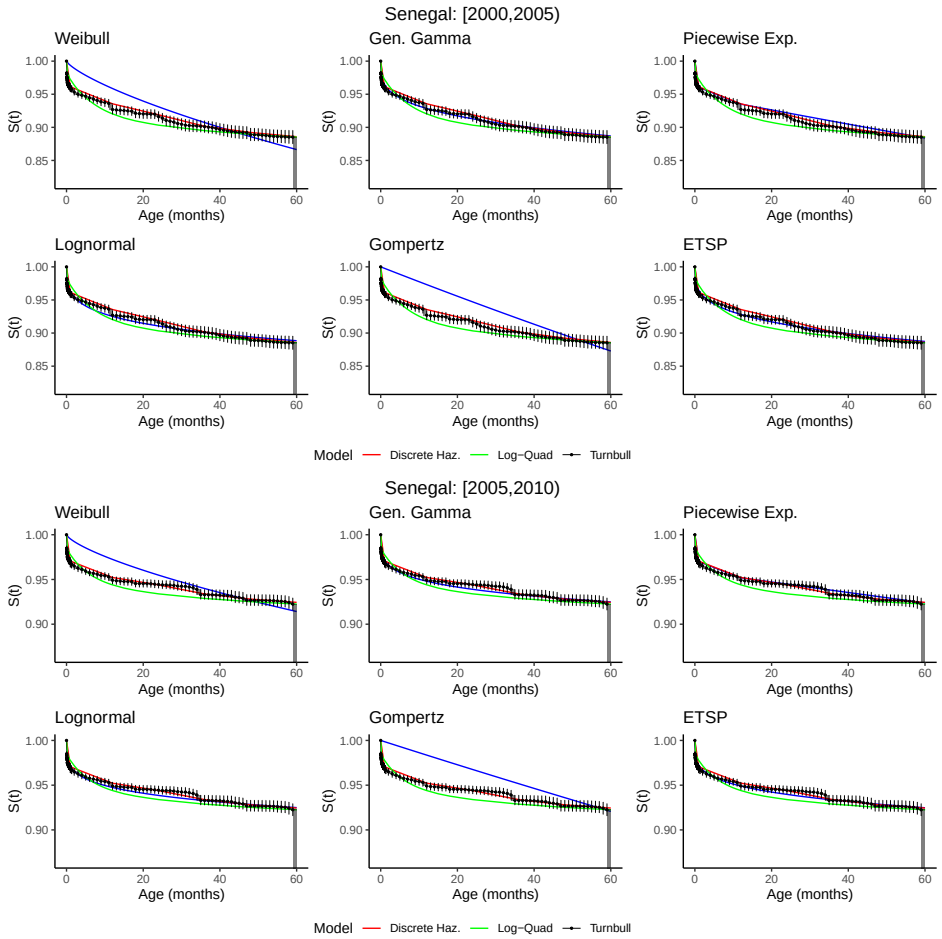
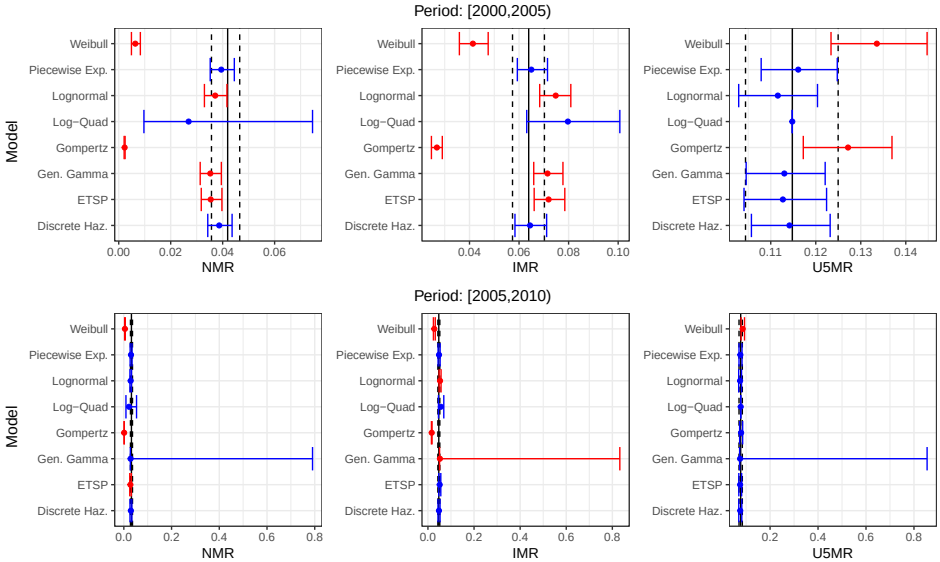


Figure A-22: Estimated survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months



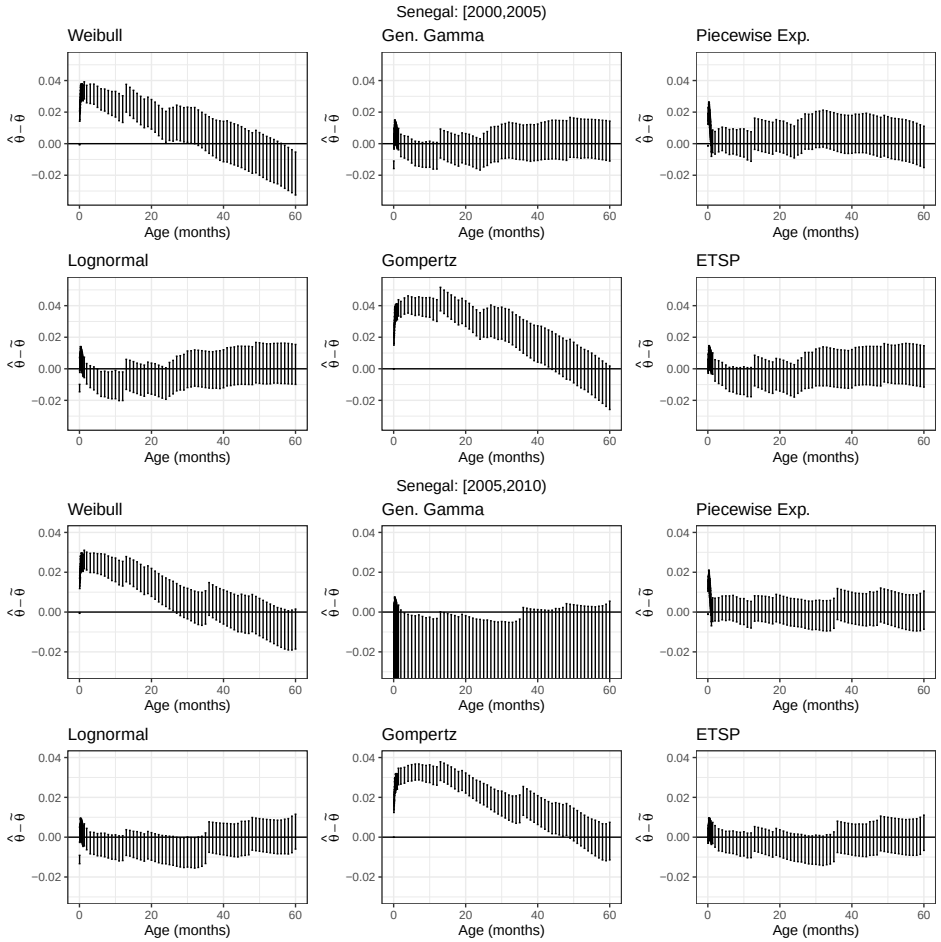
Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-23: Estimates of NMR, IMR, and U5MR for Senegal in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months



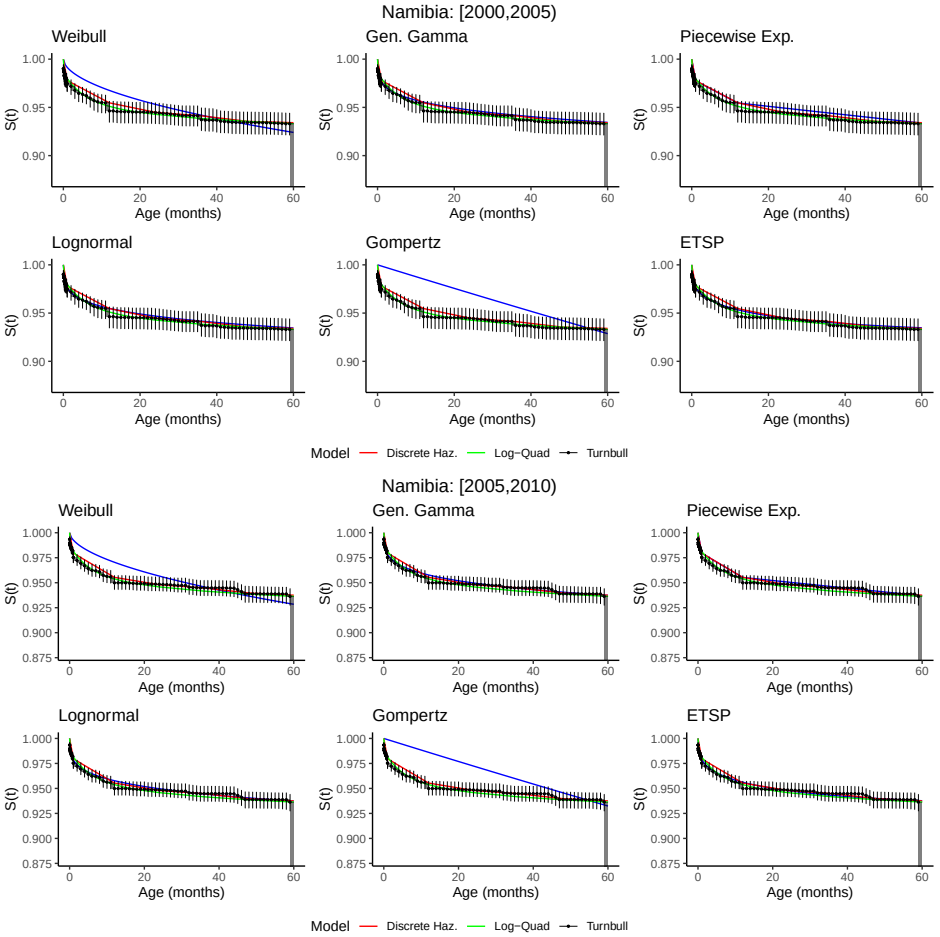
Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-24: Empirical distributions of differences in survival curves for Senegal in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$



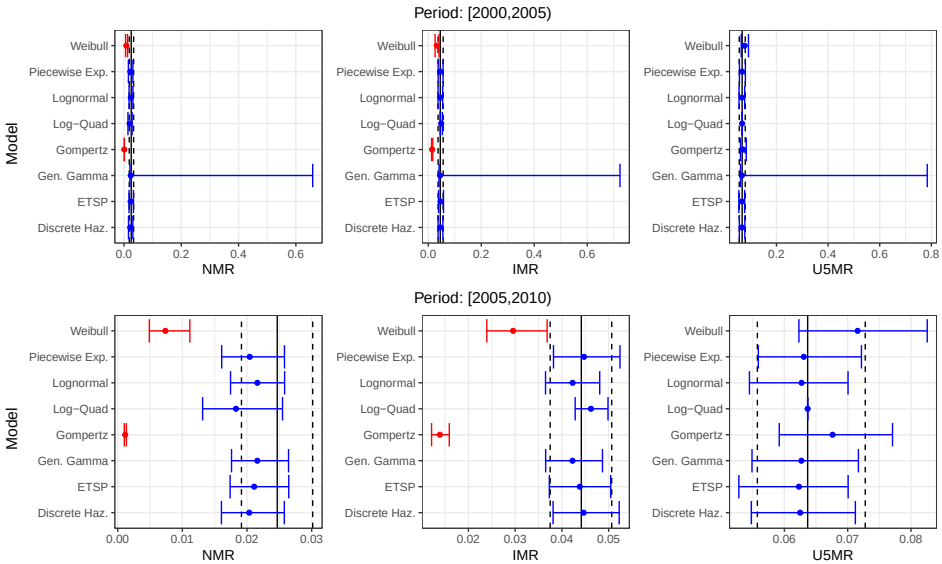
Notes: Note that for $[2005, 2010)$ the differences have been cut off at -0.03 for clarity, though the differences extend much further negative for the generalized gamma model.

Figure A-25: Estimated survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months, not adjusted for age heaping at 12 months



Notes: Parametric, pseudo-likelihood estimates are in blue. All confidence bands are 95% confidence intervals based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-26: Estimates of NMR, IMR, and U5MR for Namibia in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom), not adjusted for age heaping at 12 months



Notes: Turnbull point estimates are denoted by vertical black lines. All 95% confidence intervals are based on finite population variances, with the exception of the log-quad model where uncertainty is calculated as in Guillot et al. (2022).

Figure A-27: Empirical distributions of differences in survival curves for Namibia in $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom) from ages 0 to 60 months between parametric estimates (not adjusted for age heaping) $\hat{\theta}$ and the Turnbull estimate $\tilde{\theta}$

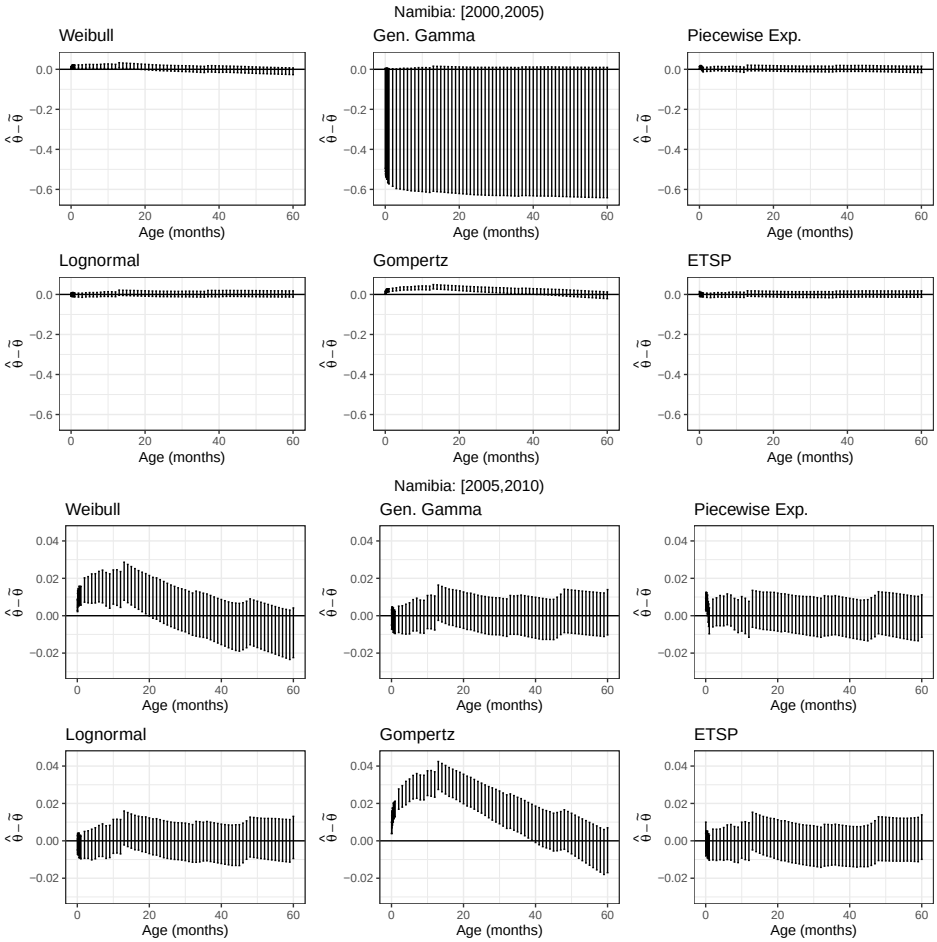


Table A-4: Model validation results

Country	Period	Weibull	Piecewise Exponential	Generalized Gamma	Lognormal	Gompertz	ETSP	Discrete Hazards
Burkina Faso	[2000, 2005)	18	37	**91**	67	15	67	64
	[2005, 2010)	27	**73**	**75**	65	12	70	**74**
Malawi	[2000, 2005)	42	66	**94**	**85**	19	**92**	**76**
	[2005, 2010)	52	**76**	**92**	**86**	17	**88**	**76**
Senegal	[2000, 2005)	29	**73**	**85**	**78**	20	**85**	**72**
	[2005, 2010)	40	**73**	62	**90**	14	**94**	**72**
Namibia	[2000, 2005)	58	**86**	**100**	**100**	27	**99**	**86**
	[2005, 2010)	56	**86**	**100**	**100**	29	**99**	**85**

Notes: Percentage of samples (out of 500) from $\hat{\theta} - \tilde{\theta}$ that contain 0 for all parametric models, countries, and periods for models that do not adjust for age heaping. Results that contain more than 70% of samples noted in bold.

In the following plots, we compare the parametric survival curves and Turnbull estimators when age heaping is adjusted for vs. unadjusted. Note that the point estimates for the survival curves are extremely similar for the Weibull, lognormal, Gompertz, generalized Gamma, and ETSP models, only differing in the third or fourth decimal place. This suggests that age heaping occurring between 6 and 18 months does not greatly impact the overall shape of the survival curve. The age-heaping-adjusted piecewise exponential model differs quite significantly from the unadjusted piecewise exponential model at age 12 months, which is to be expected based on how we interval censored the data in the adjusted model. Also note that in all cases, the uncertainty surrounding the age-heaping-adjusted model is slightly larger than the uncertainty surrounding the unadjusted models, though perhaps not meaningfully larger.

Figure A-28: Comparison of parametric models where data is adjusted for age heaping at 12 months versus not for Burkina Faso in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)

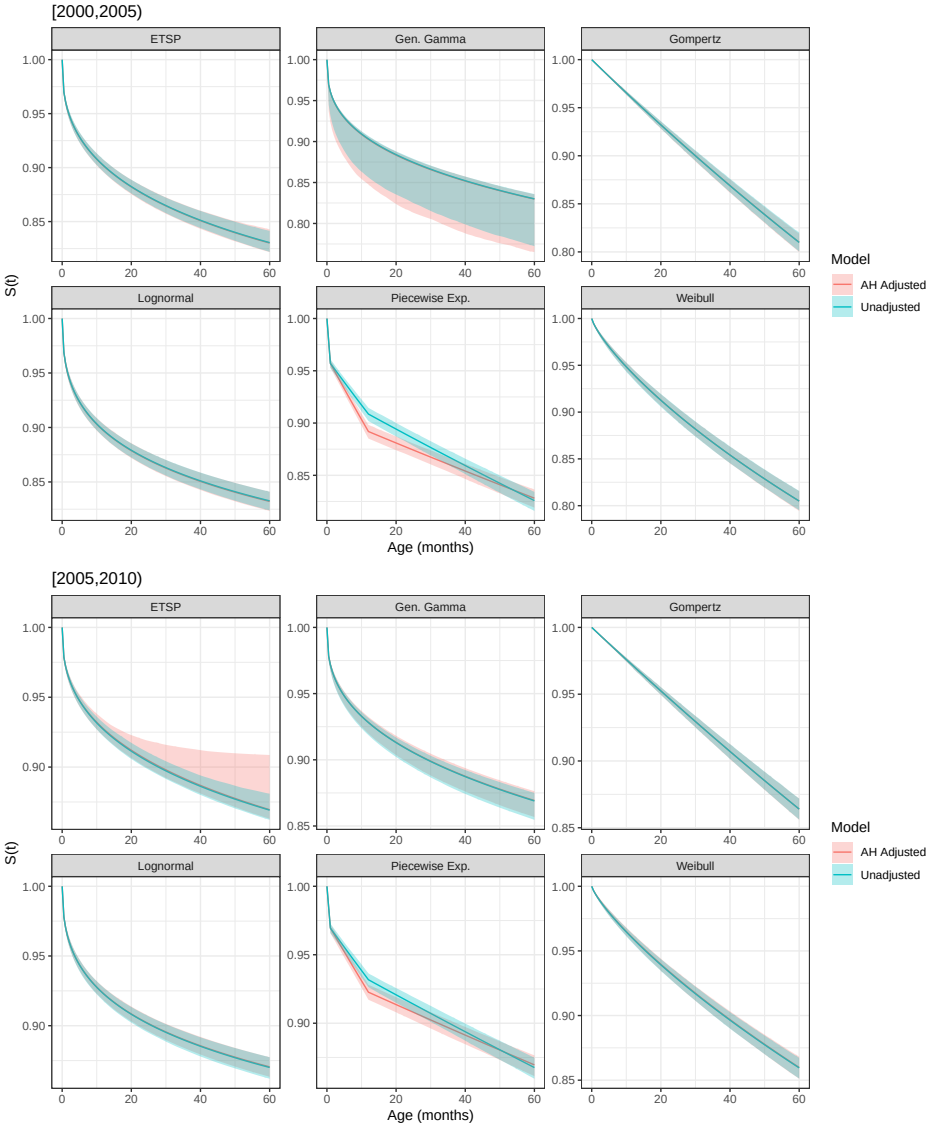


Figure A-29: Comparison of Turnbull estimator where data is adjusted for age heaping at 12 months versus not for Burkina Faso in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right)

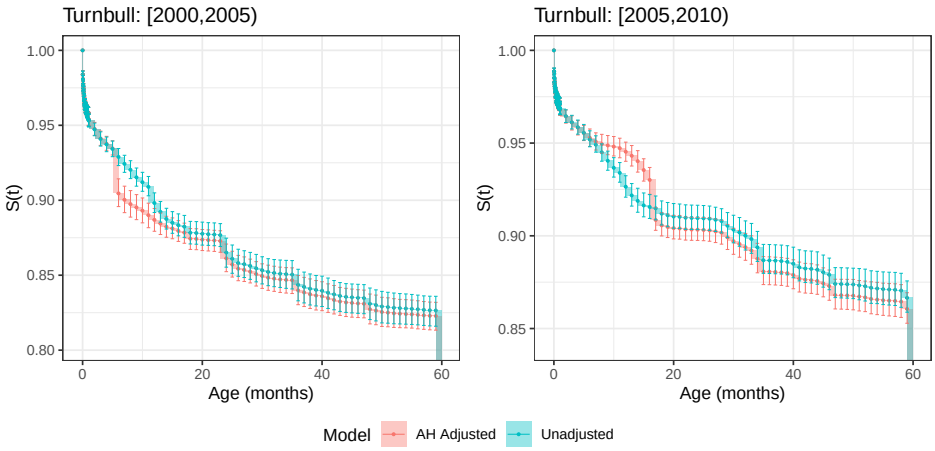


Figure A-30: Comparison of parametric models where data is adjusted for age heaping at 12 months versus not for Malawi in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)

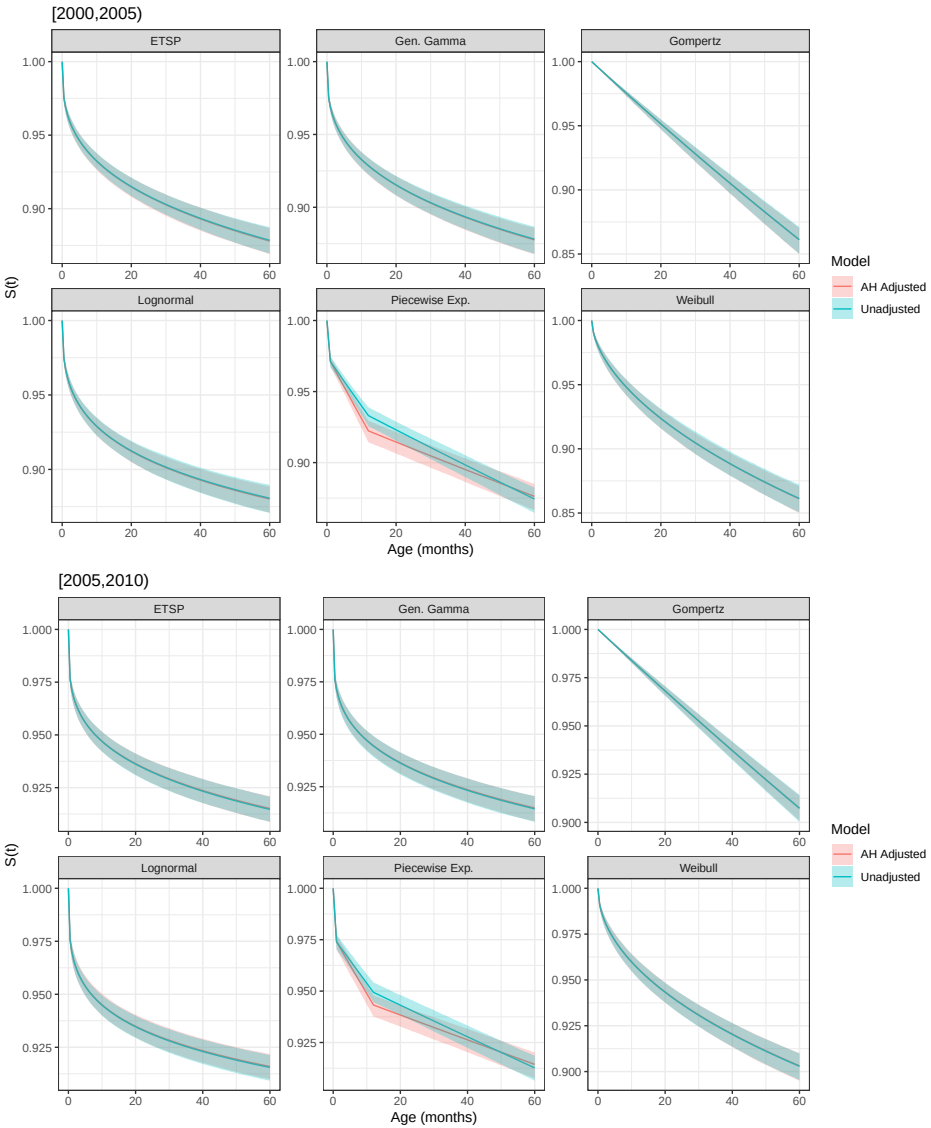


Figure A-31: Comparison of Turnbull estimator where data is adjusted for age heaping at 12 months versus not for Malawi in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right)

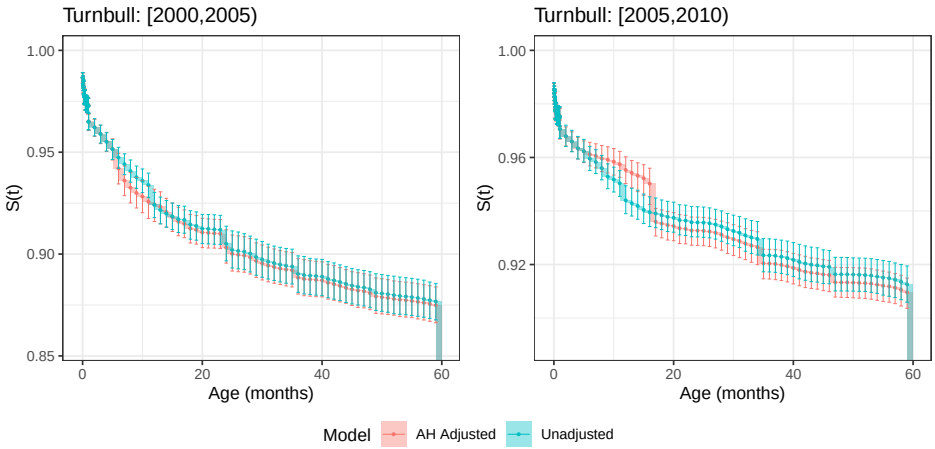


Figure A-32: Comparison of parametric models where data is adjusted for age heaping at 12 months versus not for Senegal in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)

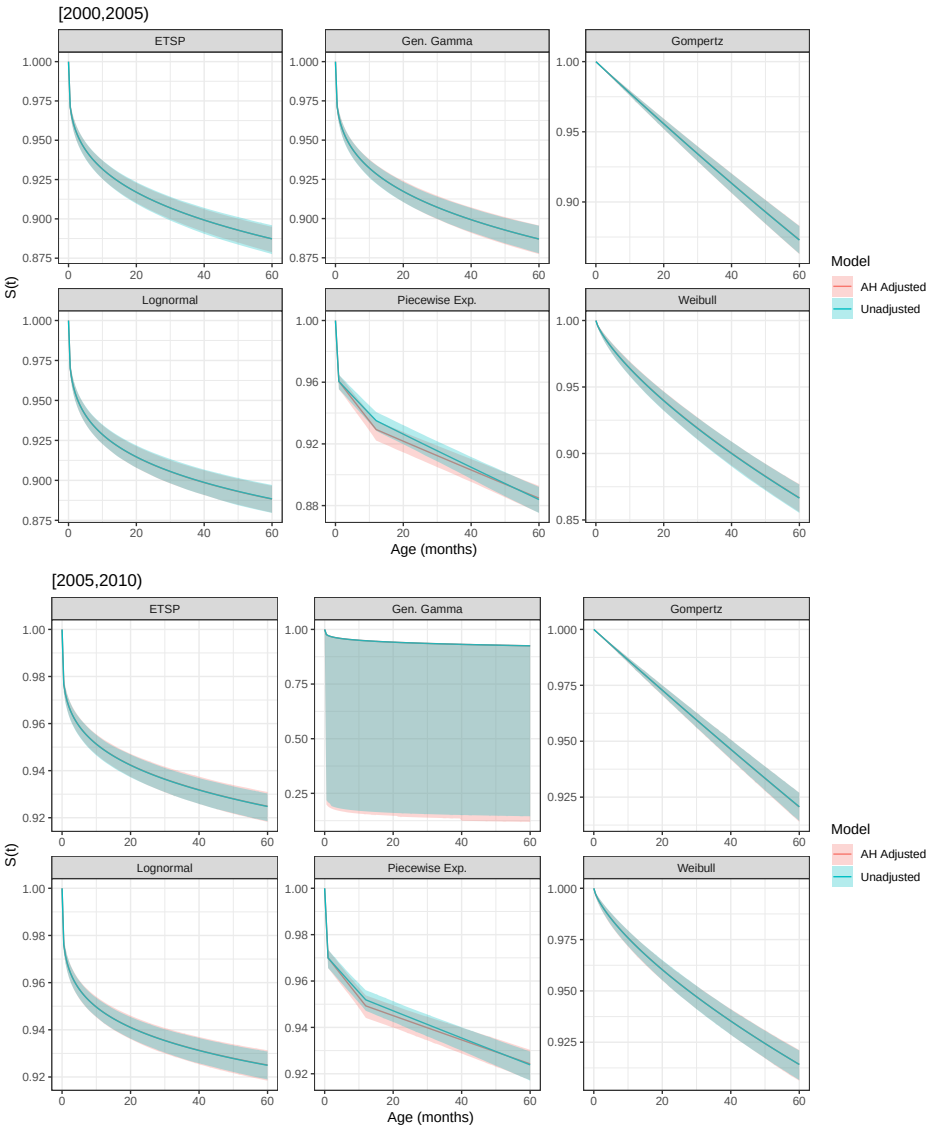


Figure A-33: Comparison of Turnbull estimator where data is adjusted for age heaping at 12 months versus not for Senegal in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right)

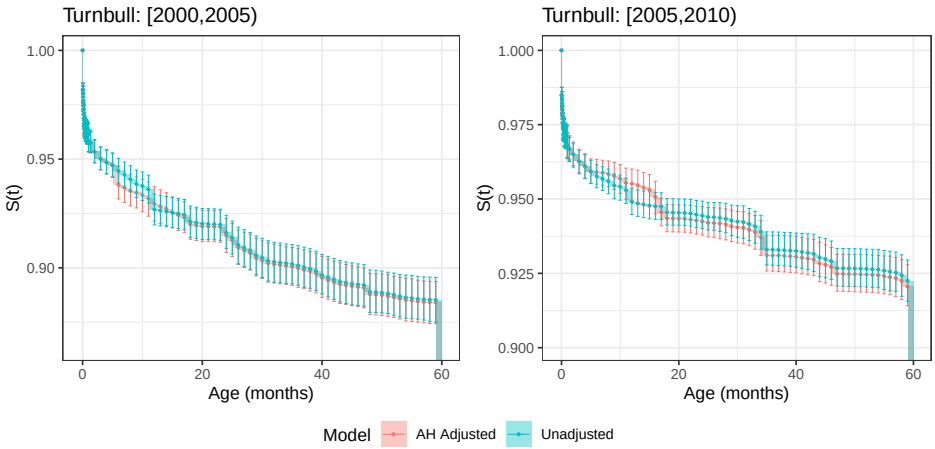


Figure A-34: Comparison of parametric models where data is adjusted for age heaping at 12 months versus not for Namibia in periods $[2000, 2005)$ (top) and $[2005, 2010)$ (bottom)

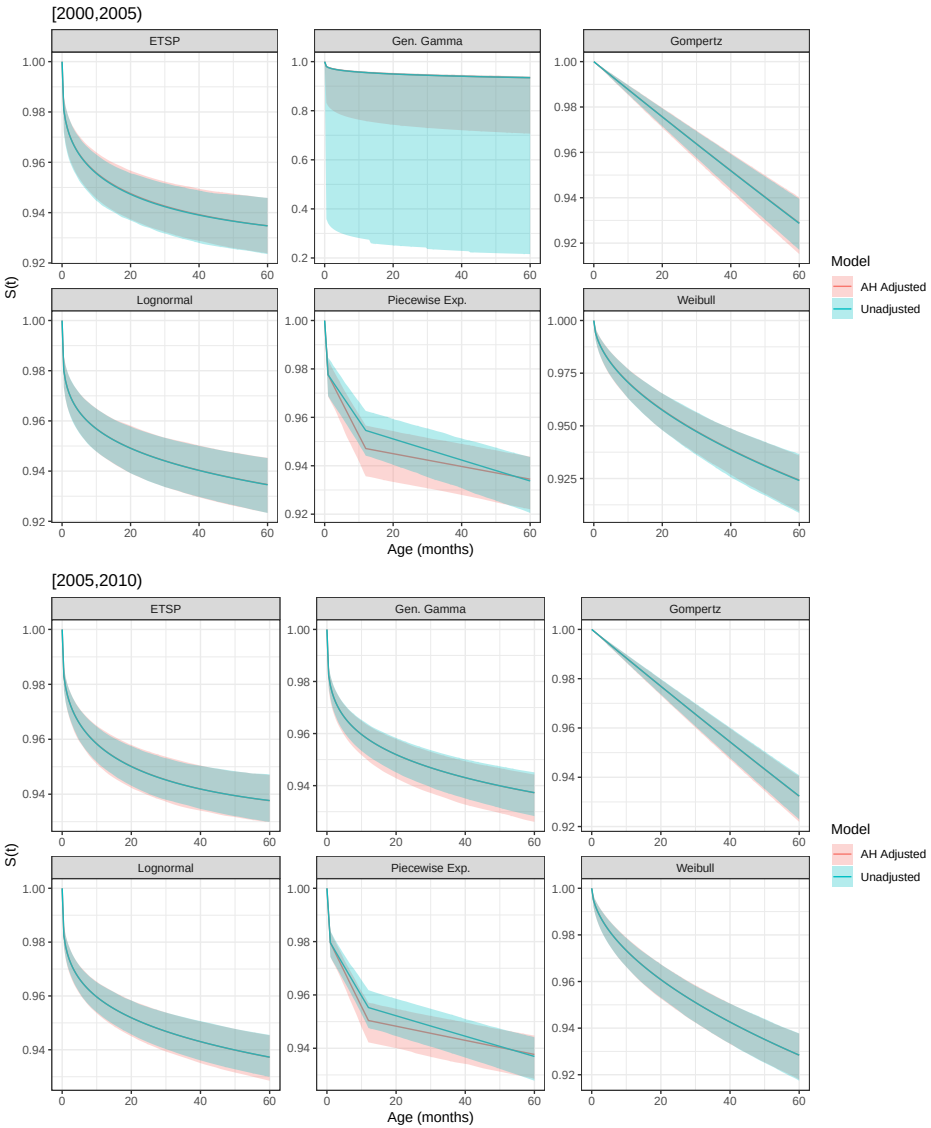


Figure A-35: Comparison of Turnbull estimator where data is adjusted for age heaping at 12 months versus not for Namibia in periods $[2000, 2005)$ (left) and $[2005, 2010)$ (right)