Correcting historical mortality rate bias in big crowd-sourced online genealogies

Michael Y.C. Chong*[1], Diego Alburez-Gutierrez[2], Monica Alexander[1,3], Emilio Zagheni[2]

[1]Department of Statistical Sciences, University of Toronto, Canada

[2]Max Planck Institute for Demographic Research, Germany

[3]Department of Sociology, University of Toronto, Canada

July 25, 2024

Abstract

In recent years, online communities have created genealogical records that span multiple continents over several centuries and contain demographic information for millions of people. However, these data are unrepresentative, and extracting accurate population information presents a major methodological challenge. We construct a Bayesian model that combines structured mortality estimation and smoothing techniques to correct mortality rates derived from FamiLinx, a large crowd-sourced genealogical dataset. Our model estimates and extrapolates a set of adjustment factors that capture the discrepancy between genealogy-derived rates and more reliable data from the Human Mortality Database. We apply our method to estimate 19th-century mortality rates for countries and time periods that are not covered by the high quality data to demonstrate out-of-sample performance. Our results illustrate a wide range of underreporting patterns across age, time, and between countries. In particular, we find that mortality is most severely underreported among young ages, ranging from a factor of 1/2 to under 1/10 the estimated true mortality rate. Understanding and accounting for these biases will be critical to future research using these data.

*myc.chong@mail.utoronto.ca

1 Introduction

Family histories have been used extensively to study historical demographic processes. Traditionally, these have taken the form of family reconstitutions (Henry, 1956; Wrigley and Schofield, 1983) and small ascendant genealogies, where individuals reconstruct their ancestry retrospectively. Increased data availability has allowed demographers to use genealogical datasets to study demographic change more broadly (Holden and Boudko, 2018; Zhao, 2001), and in recent years the internet has allowed a growing community of online genealogists to crowd-source genealogical datasets of unprecedented scale. The best example of this is FamiLinx, a large and freely accessible genealogical dataset spanning countries and centuries.

The FamiLinx dataset contains individual-level information on the places and dates of birth and death of 86 million individuals and kinship ties between parents and children (Kaplanis *et al.*, 2018). These data represent contributions from users of the website geni.com and has been curated to remove duplicates and other inconsistencies. It has an unparalleled chronological and geographic coverage, encompassing most of the Western World over the past five centuries, and crucially, unlike register-based genealogies, kin ties are not restricted by national borders, making FamiLinx a truly transnational data source.

Despite their great potential, online genealogies are subject to many biases that restrict their usability (Hollingsworth, 1976). First, in most cases, the inclusion of an individual in an ascendant genealogy is dependent on having a living descendant or relative to record them. The production of genealogies is also contingent on historical and social forces. For example, elites, who leave a longer paper trail, are more likely to be included in online genealogies (Zhao, 2001). Processes of 'selective remembering' will also affect the inclusion of individuals, whereby some ancestors are deemed more worthy of being included in a family tree than others, and indeed family histories have been found to underrepresent women, early deaths, and marriages without children, to name only a few deficiencies (Stelter and Alburez-Gutierrez, 2022). Recent work by Calderón-Bernal *et al.* (2023) uses micro-simulation to understand how certain mechanisms in the recording process, such as the omission of those without direct descendants, influence the apparent demographic rates in the genealogy.

However, the digital and crowdsourced origin of the data complicates the picture further. While some demographic characteristics of the user population (i.e. the genealogists) are available, we are unable to characterize their access to their family histories, and furthermore their use of the online platform, which could vary across different sub-populations. In spite of these serious concerns regarding data quality, existing studies using online genealogies have taken the data at face-value or assumed that the data is representative of the population at large (Blanc, 2020; Kaplanis *et al.*, 2018). Disentangling data-related processes from the population processes is a central problem in demographic research using these data. Until now, there has been no way to assess the extent and structure of biases in online genealogical data without reference data on the same population.

We propose a Bayesian model to both estimate and correct for bias in mortality rates derived from FamiLinx. We compare mortality rates derived from FamiLinx to those from the Human Mortality Database (HMD), which we treat as gold standard data for historical mortality rates (Human Mortality Database, 2024). Our results describe how the mortality rate bias varies across age, gender, countries, and time, and our method provides a mechanism to estimate mortality rates where reliable data is not available. As we will discuss in Section 3, this is made possible by a set of key assumptions regarding the smoothness of the adjustment factors and shape of mortality. We demonstrate the potential of these data as a tool to understand historical demographic trends when combined with an appropriate correction procedure.

We focus on mortality in this study for several reasons. One consideration is that historical data on mortality is more extensive than other demographic indicators. For example, the Human Fertility Database, the equivalent database for fertility, contains 19th century data for only one country as of the time of writing (Human Fertility Database, 2024), and there is no equivalent data source for migration. Additionally, the data in FamiLinx is also not well suited to study migration. Only locations of births and deaths are given, and so while one can infer the occurrence of migration events between birth, births of children, and death, the exact timing is unknown and short-term migration events may not be captured. Finally, there exist strong age patterns of human mortality as well as established methods to leverage these patterns to regularize estimates in the contexts of small area estimation (Dharamshi *et al.*, 2023; Gonzaga and Schmertmann, 2016; Schmertmann and Gonzaga, 2018) and model life tables (Clark, 2019; Wilmoth *et al.*, 2012).

Regularization to known shapes of mortality curves plays a key role in our modelling approach. The complexity of and lack of information about the data-generating process make it difficult to know how much of an adjustment is necessary for a group of interest. Therefore, instead of modelling the data-generating process explicitly, our Bayesian framework takes into account the plausibility of the adjusted mortality rates as part of the estimation of the bias. We show that this technique is capable of identifying novelties in the mortality rate bias even for populations where the reference HMD data is not available.

The rest of this paper is structured as follows. First, we introduce the two datasets used in the analysis: the FamiLinx online genealogies, and the Human Mortality Database. We describe how we processed the profiles in FamiLinx to extract age-specific mortality rates. Then, we introduce our modelling approach and apply it to mortality rates for a selection of countries. We examine the resulting bias and adjusted mortality rate estimates, and perform a set of validation exercises. We present new mortality estimates in the 19th-century United States, which suggest slightly higher life expectancy prior to the U.S. Civil War than previously thought. In Section 5, we evaluate the strengths and shortcomings of our model, discuss the implications of our results, and conclude with an agenda for future research that builds on our methodological work to answer questions of substantive interest about the historical development of human dynamics.

2 Data

2.1 Data sources

We consider profiles that come from FamiLinx, a online genealogical dataset containing over 86 million anonymized individual records with known kinship ties among 43 million individuals (Kaplanis *et al.*, 2018). The data was aggregated from hundreds of thousands of family trees created by thousands of genealogists using the social networking site Geni.com. FamiLinx is a curated version of the data which includes individuals born over the last 400 years on all continents. The site users who contributed the data mostly come from countries in the Global North, and this is reflected in the location of the recorded vital events (birth and death): 55% are located in Europe, and 30% in North America. For this study, we focus on 11 European countries that have at least some easily

accessible high-quality 19th century data from the HMD: Belgium, Denmark, Finland, France, the Netherlands, Norway, Sweden, Switzerland, and the United Kingdom. We also consider the United States as an important test case and because most profiles in FamiLinx are based in the United States.

'Gold-standard' historical demographic rates come from the Human Mortality Database, a widely used and high-quality repository of harmonized mortality data. At the time of writing, age-specific mortality data in the 19th century is available for these 11 countries[1], which we summarize in Table 1. All of these are in Europe, and a full set of data spanning the entire century is only available for Sweden. From these data we use death counts and exposure-to-death (in person-years lived) stratified by country, gender, age, and time period.

Table 1: Summary of 19th century data availability in the Human Mortality Database.

Country	Earliest year available
Belgium	1841
Denmark	1835
Finland	1878
France	1816
Netherlands	1850
Norway	1846
Sweden	1751
Switzerland	1876
United Kingdom (England and Wales)	1841

2.2 Data extraction

Obtaining country and period-specific death counts and exposure-to-death from the FamiLinx dataset is challenging for a number of reasons. First, the time and location information on individuals' vital events is often incomplete. Approximately 50% of profiles are missing their year

[1]Data for Italy is available from 1872 onward, but is omitted here because of data quality concerns. Data for Iceland is available from 1838 onward, but is omitted due to small population sizes.

of birth, and among profiles with years of birth before 1900, approximately 42% are missing the year of death. Second, while there is a two-digit country code variable, it is often missing. Other geographic information is often free-text and therefore contains errors, historical (defunct) names, and names given in the country's (non-English) language (Colasurdo and Omenti, 2024). In their original paper, Kaplanis *et al.* (2018) used a geocoding service to assign countries to the free-text descriptions in each profile. Since this step may be cost-prohibitive to some researchers, we pursue a simpler string-matching approach to country assignment. Third, there is no direct information on the place of residence of individuals over the course of their life, therefore requiring assumptions about migration to obtain country-specific quantities. We describe our treatment of these three issues below.

For this study, we only consider profiles with recorded birth and death years, imputed with the recorded baptism and burial years where available and necessary. Individuals with implausible ages at death (>110 years) are excluded from the analysis. Our model estimates rates for men and women separately, and so profiles without gender recorded are also excluded.

Birth and death locations are matched to the 10 countries of interest by searching for a set of location names in the location-relevant data fields. We use names and regular expressions for countries that are provided in the package `countrycode` (Arel-Bundock *et al.*, 2018), but for some countries make modifications to prevent false matches or to accommodate certain data entry mistakes, which we specify in Appendix A. We then count deaths and exposure-to-death for each of the countries. Death counts for a country include both those where the individual's death location was specified and successfully matched to the country, and those that are imputed from their most recent vital event. In the latter case, the individual's death location is not specified, but their most recent vital event before death (either the last birth of a child, or their own birth if they do not have recorded children) does have a location specified and successfully matched to the country.

Individuals' contributions to exposure-to-death are divided among the countries in which they had vital events (birth, birth of a child, or death) according to the locations of the vital events. We think of each individual's life course as a sequence of segments between the vital events with possibly missing or different location endpoints. If both endpoints are missing, that segment is not considered. Once every segment has at least one specified location endpoint, they contribute to a country's

exposure-to-death if at least one of the endpoints matches the country.

To illustrate, consider an individual with three recorded vital events: (1) born in Sweden, (2) had a child in Sweden at age 25, and (3) died in the United States at age 90. We can distinguish two life segments: from age 0 to 25, and 25-90. All 90 years lived would be included in the exposure-to-death for Sweden, and the segment from age 25-90 would also be counted towards the exposure-to-death for the United States. A diagram illustrating examples of calculation of deaths and exposure-to-death from profiles is given in Figure 1.

Figure 1: Diagram illustrating death and exposure calculations from vital events locations in FamiLinx profiles. The life courses of seven hypothetical profiles are represented by segmented vertical timelines, followed by circles representing their deaths. Recorded locations of vital events are shown along each timeline (A, B, C, or missing), where only A and B are countries of interest. The colours of the segments represent whether they are counted in the exposure-to-death of the countries of interest, and the colours of the circles represent where their death is counted.

For most of the populations, the contribution of the "migrant" part of the exposure and imputed portion of deaths are relatively small, but there are exceptions. For example, migration from Great Britain appears high in this period, whereas migration from the Netherlands is relatively very small, as shown in Figure 2. The full set of plots showing the composition of the death counts and exposure-to-death is shown in Appendix B, and we discuss the implications of these assumptions in more detail in Section 5.

Figure 2: Composition of exposure-to-death in Great Britain and the Netherlands in the 1850-1855 period. For each age group, the exposure-to-death is divided to show the contributions of different kinds of life course segments: "did not move" represents segments that have both endpoints in the country of interest, "missing information" represents segments that have one endpoint with in the country of interest and the other is missing, "emigrants" represent segments with a starting point in the country of interest but end in a different country, and "immigrants" represent the segments with a starting point in a different country but end in the country of interest.

Due to suspected data quality issues, we omit data from the Netherlands before 1820, where there is a uniquely sharp drop in the implied mortality rate, particularly in the youngest ages. In these periods, the region was undergoing frequent political change, which could have resulted in limited access to family histories, rapid demographic change, and inconsistent location naming, which we are unable to account for with our extraction strategy.

3 Statistical model

3.1 Overview

We use a Bayesian modelling framework to jointly estimate the discrepancy between the FamiLinx mortality and the HMD. Our model treats death counts from both of the data sources as observations from the same true underlying mortality rate, but adds an adjustment factor for the FamiLinx data. While in cases where HMD data is available, the adjustment factor is readily apparent as the ratio of mortality rates between the two data sources, extrapolating to where HMD data is not available presents a dilemma. On one hand, in order to properly identify the mortality rate from FamiLinx data, the adjustment factor cannot be left free to vary; otherwise the observation can be explained by either the mortality rate or adjustment factor parameter. On the other hand, we have to have enough flexibility to accommodate variation in the adjustment rate across contexts, because the level of bias could depend on the country, age, time period, and the interaction between these variables.

We balance these competing considerations in our framework by combining three key elements. First, we adapt ideas from Alexander *et al.* (2017) to formulate a mortality model that enforces certain shapes of the mortality curve over age. Second, we accommodate shocks and deviations from the structured mortality model using sparsity-inducing priors. In particular, by using the regularized horseshoe prior of Piironen and Vehtari (2017), we are able to control the magnitude and sparsity of the deviations even when they are weakly identified. Finally, our model performs 2D-smoothing of the adjustment factor over age and time. We use tensor product smoothing (Wood *et al.*, 2013; Wood, 2017) with a hierarchical approach to regularize the adjustment factor surfaces.

Under this setup, isolated mortality shocks can be explained by the sparse deviation terms, while persistent variation away from expected mortality shapes are explained mostly by the adjustment factor. A schematic diagram of the model is given in Figure 3 and we describe these components of the model in more detail below.

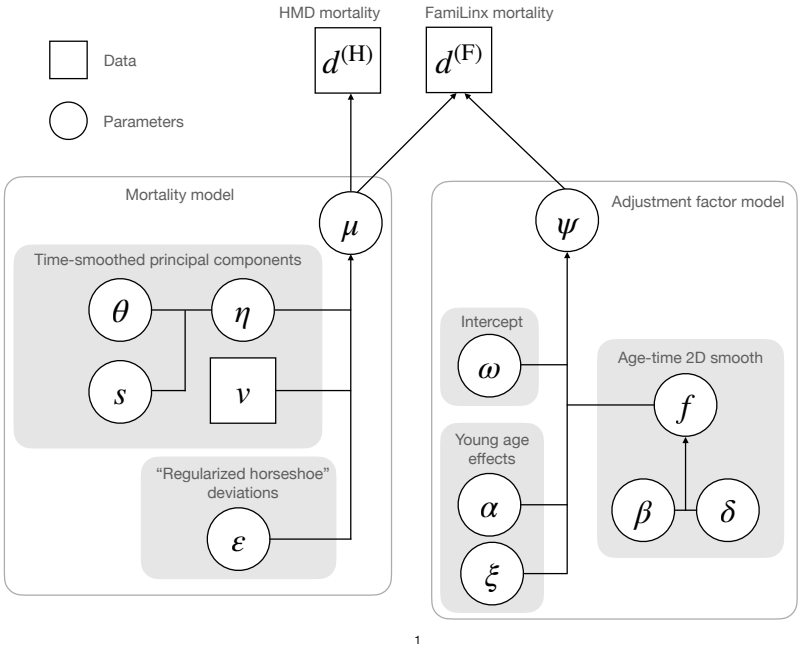


Figure 3: Schematic diagram of the estimation model.

3.2 Data likelihood

Let $d^{(\mathrm{H})}_{c,g,t,x}$ and $P^{(\mathrm{H})}_{c,g,t,x}$ denote the HMD death counts and exposure-to-death (in person-years lived) for country c, gender group g, time period t, and age group x, and let $d^{(\mathrm{F})}_{c,g,t,x}$ and $P^{(\mathrm{F})}_{c,g,t,x}$ denote the analogous extracted quantities from FamiLinx as described in Section 2. We model the death counts as coming from Negative Binomial distributions, which better account for noisiness in death counts. For the HMD counts, we have

$$d^{(\mathrm{H})}_{c,g,t,x} | \mu_{c,g,t,x}, \phi^{(\mathrm{H})}_{x} \sim \mathrm{NegBinom}(\mu_{c,g,t,x} P^{(\mathrm{H})}_{c,g,t,x},\, \phi^{(\mathrm{H})}_{x})$$

and for the FamiLinx counts we have

$$d^{(\mathrm{F})}_{c,g,t,x} | \mu_{c,g,t,x}, \phi^{(\mathrm{H})}_{x}, \psi_{c,g,t,x} \sim \mathrm{NegBinom}(\mu_{c,g,t,x} \psi_{c,g,t,x} P^{(\mathrm{F})}_{c,g,t,x},\, \phi^{(\mathrm{F})}_{x}),$$

where $\phi^{(\mathrm{H})}_{x}$ and $\phi^{(\mathrm{F})}_{x}$ are age-specific overdispersion parameters, and $\psi_{c,g,t,x}$ is an adjustment factor for the FamiLinx data.

3.2.1 Adjustment factor

We first discuss the model for the adjustment factor. Given the lack of prior knowledge about the adjustment factor, our goal for this component is to allow adequate flexibility to capture patterns, possibly across age, time, gender, and country. We therefore construct the adjustment factor as the sum of a two-dimensional tensor product spline smooth f, an additional term α for age 0, and an additional term ρ for age groups younger than 20 such that

$$\log(\psi_{c,g,t,x}) = \omega_c + \alpha_c I(x = 0) + \xi_c I(x < 20) + f_{c,g}(x, t).$$

Here, $f_{c,g}(t, x)$ represents a country- and gender-specific smooth 2D function over time periods and age groups constructed as a tensor product smooth interaction (Wood, 2017). This allows for the age patterns to vary across time, and accommodates for populations from different countries having different structures and extents of mortality rate bias.

We explain our setup and use of the tensor smooth below. For greater technical detail about the parameterization we use, refer to more comprehensive explanations given by Wood *et al.* (2013) and Pedersen *et al.* (2019). Intuitively, the construction of f can be thought of as

$$f(x, t) = \sum_{kj} \beta_{kj} h_j(t) g_k(x),$$

where the sets of h_j and g_k represent basis spline functions for time and age respectively, and β_{kj} represent coefficients. We use cubic regression spline bases for both, with a 2nd derivative marginal penalties. For the marginal spline basis over age, we use 10 knots, placed at ages 1, 3, 5, 7.5, 10, 20, 30, 50, 70, and 80. For the marginal spline basis over time, we place 9 equally spaced knots from year 1800 to 1900. We use the parameterization given by Wood *et al.* (2013), whereby $\vec{f} = [f(x_1, t_1), \ldots, f(x_N, t_N)]^T$ can be represented as a mixed model

$$\vec{f} = \mathbf{X}\vec{\beta} + \mathbf{Z}\vec{\delta},$$

where $\mathbf{X}$ represents unpenalized fixed effects with associated coefficients $\vec{\beta}$ and $\mathbf{Z}$ represents penalized components with associated random effects $\vec{\delta}$. Using a typical 2nd derivative penalty, X

contains the linear and constant functions, since these have a 2nd derivative of zero. In our hierarchical fully Bayesian approach, we have that

$$\vec{f}_{c,g} = X\vec{\beta}_{c,g} + Z\vec{\delta}_{c,g}. \tag{1}$$

The fixed effect slopes $\beta_{c,g}$ are modelled

$$\vec{\beta}_{c,g} = \vec{\beta}_g + \vec{\gamma}_{c,g},$$

where $\vec{\beta}_g$ is a vector of global gender-specific coefficient means given i.i.d. standard normal priors,

$$\vec{\beta}_g \sim MVN(\vec{0}, I_{3\times 3}).$$

The $\vec{\gamma}_{c,g}$ terms represent countries' variations around the slopes, and are given priors

$$\vec{\gamma}_{c,g}|\vec{\sigma}_\gamma \sim MVN(0, \text{diag}(\vec{\sigma}_\gamma^2))$$

where each component of $\vec{\sigma}_\gamma^2$ is given a $N^+(0,1)$ prior.

The term relating to Z is also defined hierarchically, where

$$\vec{\delta}_{c,g} = \vec{\delta}_g + \vec{\zeta}_{c,g},$$

where δ_g determines a global nonlinear trend for each gender, and each $\delta_{c,g}$ represents country-specific variation from those global trends. Priors are similar to those above on β and γ, but are slightly more involved. We describe this in more detail in Appendix E.

The hierarchical treatment of the entire spline $f_{c,g}$ is clearer if we rewrite Equation 2 as

$$\vec{f}_{c,g} = \left(X\vec{\beta}_g + Z\vec{\delta}_g\right) + \left(X\vec{\gamma}_{c,g} + Z\vec{\zeta}_{c,g}\right)$$

such that the first two terms capture global trends in the adjustment factor over age and time, and the last two terms capture country-specific variation away from that trend.

The intercept ω_c represents a country-level average adjustment factor over age and time, around which f fluctuates. The age 0 and "young age" effects α_c and ξ_c are country-specific adjustments to certain age groups. These are to capture additional volatility in the younger age groups that is difficult to reflect in the smooth function. All these are given hierarchical priors,

$$\omega_c|\omega_0, \sigma_\omega \sim N(\omega_0, \sigma_\omega^2)$$
$$\alpha_c|\alpha_0, \sigma_\alpha \sim N(\alpha_0, \sigma_\alpha^2)$$
$$\xi_c|\xi_0, \sigma_\xi \sim N(\xi_0, \sigma_\xi^2)$$

where the means and standard deviations are given weakly informative $N(0, 1^2)$ priors and $N^+(0, 1^2)$ respectively.

3.2.2 Mortality model

We now turn to the model on the the mortality curves over age, $\vec{\mu}$. These curves are expected to conform to regular shapes, which are extracted from the HMD through singular value decomposition (SVD). Specifically, let M be a N_M by A matrix of logged mortality rates, where each of the N_M rows represents the population in some country-period, A is the number of age groups. For this application, to construct M we use the set of HMD mortality rates $(d^{(H)}, P^{(H)})$ used to fit the model. We discuss the choice of M in more detail in Section 5.2 and Section 5.3.

If UDV^T represents the SVD of M, then V is the $A \times A$ matrix of right singular vectors. Constructed this way, the columns $\vec{v}_1, \ldots, \vec{v}_A$ of V represent the principal axes of the data over the age groups in order of decreasing variance explained. We can therefore express regular age patterns of mortality using only the first few columns of V.[2] To capture historical shocks in mortality and other atypical patterns, we include an additional error term.

We perform separate SVDs $M_g = U_g D_g V_g^T$, for the male and female populations. Then for country c, gender g, and time period t, we model the true age specific mortality rates $\vec{\mu}_{c,g,t} =$

[2]Empirically in other contexts, the first three or four principal axes have been enough to describe most of the structural mortality rate variation (Alexander *et al.*, 2017; Dharamshi *et al.*, 2023).

$(\mu_{c,g,t,1}, \dots, \mu_{c,g,t,X})^T$ as

$$\log \vec{\mu}_{c,g,t} = \eta_{1,c,g,t}\vec{v}_{g,1} + \eta_{2,c,g,t}\vec{v}_{g,2} + \eta_{3,c,g,t}\vec{v}_{g,3} + \eta_{4,c,g,t}\vec{v}_{g,4} + \vec{\varepsilon}_{c,g,t},$$

where $\vec{v}_{g,1}, \dots, \vec{v}g, 4$ are the first four columns of V_g, and we call $\{\eta_{k,c,g,t}\}_{k=1,\dots,4}$ the 'principal component (PC) parameters', and $\vec{\varepsilon}_{c,g,t} = (\varepsilon_1, \dots, \varepsilon_A)_{c,g,t}$ represents deviations that cannot be represented by the first 4 principal axes.

Each series of PC parameters is assumed to change smoothly over time. For each (k, c, g), we model $\eta_{j,c,g}$ as

$$\eta_{j,c,g,t} = \theta_{j,c,g} + s_{j,c,g}(t),$$

where $\theta_{j,c,g}$ is an intercept and $s_{j,c,g}(t)$ is a spline-based smooth function of t. To construct s, we assume a cubic regression spline basis, and estimate using a mixed model parameterization (described e.g. in Wood (2004, Appendix) or Wood (2017)). Note that the more involved construction from Wood *et al.* (2013) is not necessary in this case because we are not using tensor product splines. We therefore estimate s as

$$s = W\nu + R\rho,$$

where $W\nu$ represents the unpenalized component and $R\rho$ represents the penalized component.

Most of the systematic variation in the mortality rates is expected to be captured by the PC parameters. However, there may be historical shocks to mortality and other patterns in countries that are not represented subset of principal axes of M. We therefore allow for deviation from the principal axes using the error terms $\vec{\varepsilon}_{c,g,t}$, but since we believe that many of these should be close to zero, we impose a regularized horseshoe prior introduced by Piironen and Vehtari (2017) to encourage sparsity. One advantage of the regularized horseshoe over the usual horseshoe is better identification of weakly identified parameters. This is useful in our case since there are several model components competing to explain the FamiLinx mortality rates.

The regularized horseshoe prior is given by

$$\varepsilon_i|\tau,\lambda_i \sim N(0,\tau^2\tilde{\lambda}_i^2),$$
$$\tilde{\lambda}_i^2 = \frac{d^2\lambda_i^2}{d^2+\tau^2\lambda_i^2},$$

where, similar to the usual horseshoe prior, τ^2 controls global shrinkage towards zero, and $\tilde{\lambda}_i^2$ controls whether individual terms escape zero. On the hyperparameters, we place the following priors:

$$\tau \sim \text{Cauchy}(0,\tau_0)$$
$$\lambda_i^2 \sim \text{Cauchy}(0,1) \quad \text{for all } i$$

where τ_0 is set to be 0.01. Roughly speaking, τ and τ_0 control the sparsity of the ε_is. In their paper, Piironen and Vehtari (2017) demonstrate how to choose τ_0 in relatively simple models based on prior information about the number of nonzero parameters, but in more complicated models this information is not as easily translated. We check the sensitivity of the set of ε_is to the choice of τ_0 in Appendix D.

3.3 Validation exercises

To evaluate the performance of the model, we considered validation exercises in which subsets of the data were left out. For each of the countries in Table 1, we fit the model leaving out all but the latest period of HMD data (to assess performance in back-projection) and the entire HMD series (to assess performance in out-of-sample countries).

3.4 Computation

We fit the model using Hamiltonian Monte Carlo as implemented in `cmdstanr` Gabry *et al.* (2023). Obtaining 2000 samples (the first 1000 of which are discarded as warmup) per chain from 4 chains in parallel on 2.1 GHz CPUs takes approximately 3 hours. We use functions from the `mgcv` package (Wood, 2004) in R (R Core Team, 2022) to set up the spline design matrices. Computing was enabled in part by support provided by Compute Ontario (https://www.computeontario.ca/) and the Digital Research Alliance of Canada (https://alliancecan.ca).

Code to prepare the data and reproduce the results in this paper are given in the GitHub repository: https://github.com/michael-chong/familinx-mort.

4 Results

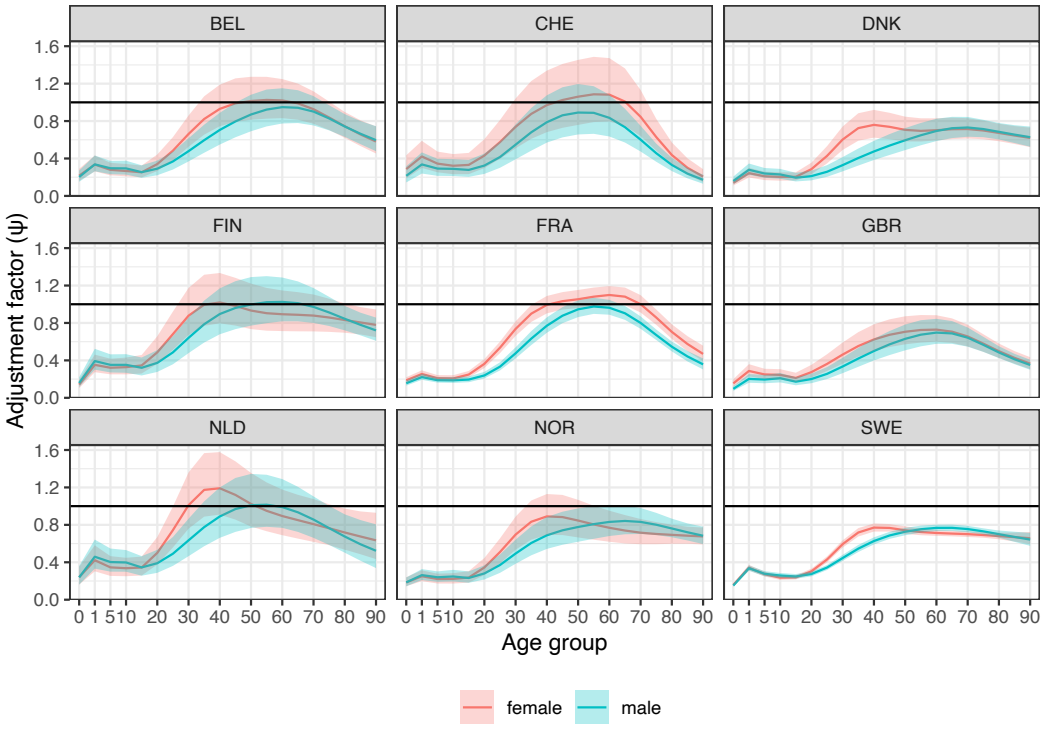
4.1 Adjustment factor estimates

We first explore estimates of the estimated adjustment factor ψ. In Figure 4a, we show estimates of the adjustment factor for the period 1800-1805 for the set of countries of interest, and in Figure 4b we show estimates for the period 1895-1899. We select these periods to contrast between the beginning and end of the study period. Our model captures substantial variation across age, time, and geography in the adjustment factors, and we highlight key patterns here.

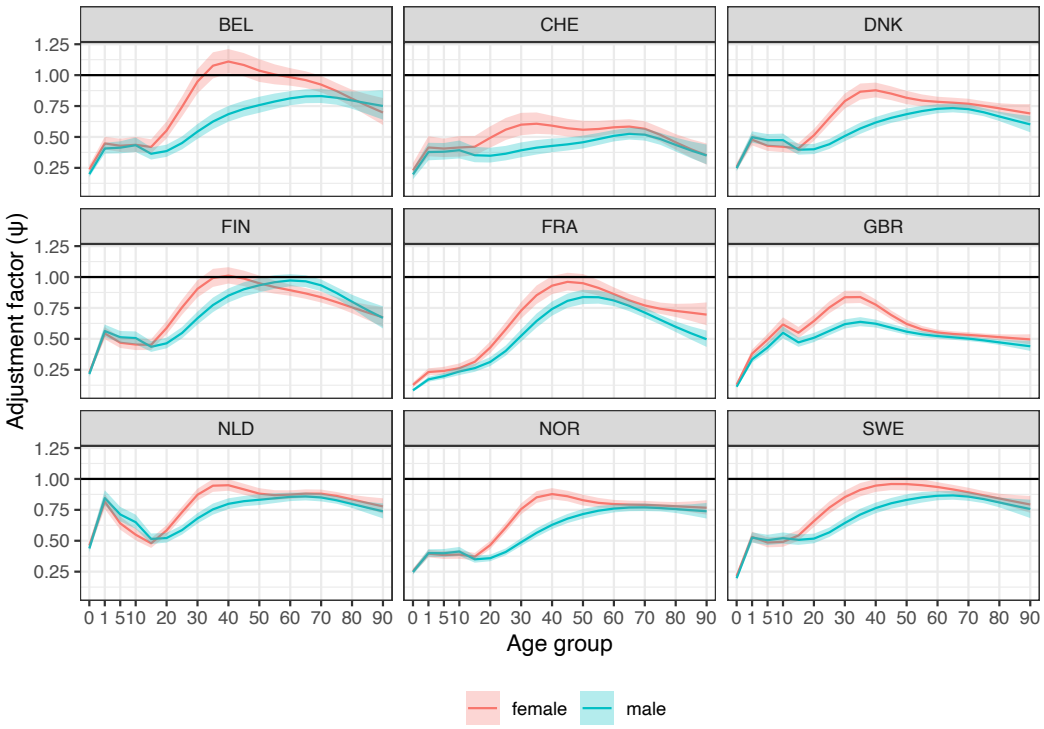
Based on our existing understanding of biases in other genealogies, we expect deaths in early life to be underrepresented. Indeed, we find that, overall, infant mortality is severely underestimated from the FamiLinx data. The estimated adjustment factor in the 0 to 1 age group in 1800 Great Britain, for example, indicates that the mortality rate inferred from the FamiLinx profiles is around 4 times lower than the true mortality rate for this group. At the other extreme, the mid- and late-adult mortality rates inferred from the genealogy are closer to the gold-standard data.

However, the shape of the adjustment curve varies greatly. In France, for example, the shape over age is parabolic, with relatively more underreporting of mortality at the youngest and oldest ages, while in Norway, the adjustment curve is relatively flat after age 40. Perhaps the most interesting age groups are those spanning childhood to early adulthood, which exhibit a range of different shapes. In the Netherlands for example, the estimated adjustment factor dips in the early adult ages, in Sweden the curve is relatively flat in this range, and in France the adjustment factor increases with age.

The shapes of the curves can vary over time. In Figure 5, we show the estimated adjustment factor surface over age and time for populations in Sweden and Great Britain. For Sweden, the curves remain relatively stable over time, particularly for the adult age group, and the shape of the curve over age remains consistent. For Great Britain however, the shape over age changes substantially

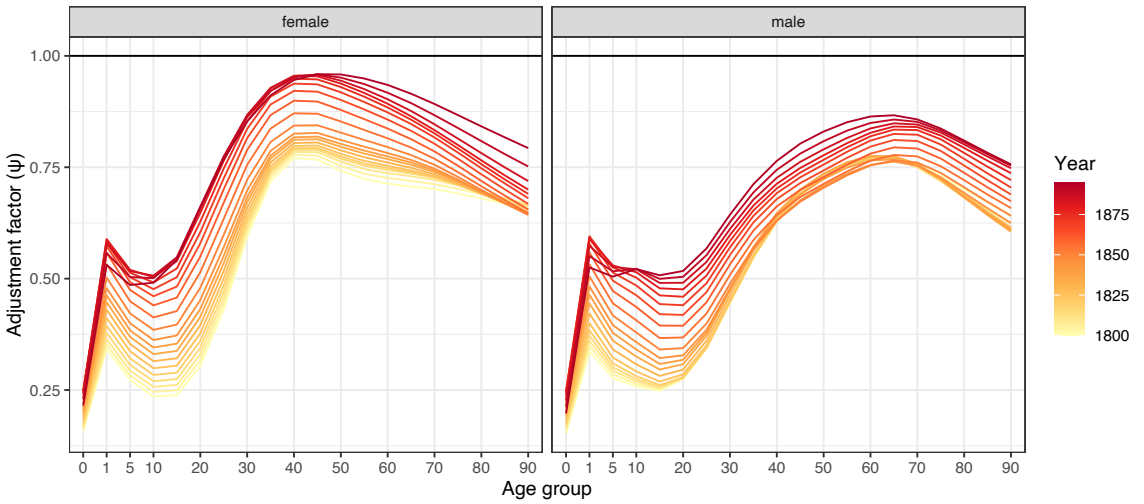


(a) Years 1800-1805.



(b) Years 1895-1899.

Figure 4: Adjustment factor for European countries in 1800 and 1895. Lines represent posterior medians, and the shaded regions represent 90% credible intervals.



(a) Sweden.

(b) Great Britain.

Figure 5: Adjustment factor curves for Sweden and Great Britain in the 19th century. Lines represent the posterior median estimate for each time period.

over the period of interest, meaning that the relative degrees of mortality underreporting between age groups changes over time. In the early part of the century, we estimate that adolescent and young adult mortality is severely underreported, but by the end of century the bias is comparable to that in the older adult ages. This may partially be due asymmetric impacts of assumptions regarding migration. We revisit the implications of these differences in Section 5.

The United States is a special case among the countries of interest because of the concentration of the userbase here, its context as the only non-European country, and also because we do not have any HMD data in this period. We therefore suspected there to be differences between the U.S. and the other countries, but the lack of data means that the patterns estimated for the United States are not informed by comparison to high quality data, but rather a mix of pooled information from the other countries and regularization on the shape of the mortality curve. We show the adjustment factor estimates for the United States in Figure 6, which exhibit different patterns than those of the other countries. We estimate that early age mortality underreporting in the FamiLinx dataset is most severe in the U.S., but is comparable to the other countries by around age 35.

Figure 6: Adjustment factor curves for the United States in the 19th century. Lines represent posterior median estimates for each time period.

4.2 Performance in left-out time periods

In Figure 7 and Figure 8, we show the mortality estimates when we leave out early-period data for select countries. For brevity we show only a few illustrative cases here: Sweden and France because these have the longest HMD series for comparison. Analogous results for other countries are similar, and are shown in Appendix C.

Figure 7: Mortality curves for Sweden from the back-projection validation exercise. Lines represent the posterior median estimate for each time period.

Figure 8: Mortality curves for France from the back-projection validation exercise. Lines represent the posterior median estimate for each time period.

In general, our estimates of the mortality curve improve on the unadjusted rates. The shapes are better aligned with the true curves, even when the FamiLinx signal is noisy, as is the case in the French population. However, the model can sometimes struggle to recover the exact true structure for earlier periods. In the case of Sweden, the model slightly overestimates early-life mortality while underestimating mid-to-late adult mortality around the middle of the century.

4.3 Performance in left-out countries

With some exceptions, the model is able to accurately recover most of the historical mortality rates when all of the data for a country is left out. We illustrate one interesting exception here, and full results are given in Appendix C. In Figure 9 we compare the estimated adjustment factors for Great Britain under three data scenarios: (1) include none of the HMD data, (2) include only the 1895-1899 data, and (3) include all of the HMD data (1841-1899). In the top row, we notice that without any HMD data, the parsimonious explanation for the FamiLinx data is a relatively stable adjustment factor curve over time. In the middle row, the adjustment factors are estimated to change more dramatically over time, and is much closer to what is estimated given the full HMD series.

Figure 9: Adjustment factor curves for Great Britain in the 19th century from the out-of-sample validation exercise. Lines represent the posterior median estimates for each time period.

This difference in performance from including no data to a minimal amount of data suggests that incorporating other types of evidence into the model, even if partial or incomplete, can help with identification of the parameters and improve the reliability of the model. We discuss possible avenues for refining estimates in Section 5.2.

4.4 Mortality rate estimates for the United States

For the United States, in Figure 10 we compare against life expectancy estimates that are compiled in Haines (2001) and Hacker (2010). Our goal with this comparison is to check whether the resulting mortality estimates are reasonable, rather than treat these data as the target. Most of the available data in this period pertain to subpopulations of the United States stratified by geography, race, or both. Two notable exceptions are the series from Haines (1979) and Hacker (2010), which attempt to estimate national mortality in the United States using model life table approaches.

Figure 10: Life expectancy estimates for the United States. The lines and shaded regions represent posterior medians and 90% credible intervals respectively. Points represent past estimates from multiple sources (see Haines (2001)), and are plotted at the midpoint of their relevant period.

Our estimates of life expectancies at age 0, 10, and 20 align roughly with the respective clusters of points in the latter part of the century. The comparisons of life expectancies at multiple ages is useful as a check on the shape of the mortality curve. That all of them seem well-calibrated implies that the segments of the mortality curve (between ages 0, 10, 20, and end of life) are individually well-calibrated also. For the earlier part of the century, our estimates show higher life expectancy (lower mortality) than Hacker (2010), and slightly different time trends. We compare our approach with that of Hacker (2010) in more detail in Section 5.3.

We are able to detect shocks in mortality that are not as evident in previous estimates. The U.S. Civil War, for example, took place in the years 1861-1865. Figure 11 shows the estimated mortality curves for 1860-1864, 1865-1869, and the neighbouring time periods, which shows a sharp increase in male mortality for the periods overlapping with the war. We also show the corresponding $\vec{\varepsilon}$ terms, which capture these deviations. The sparsity-inducing regularized horseshoe prior keeps most terms close to zero, while the terms for the affected populations are able to reflect the shock.

Figure 11: Mortality rate and deviation parameter estimates in the United States Civil War period. Lines show posterior median estimates of the corresponding quantities.

5 Discussion

In this study, we demonstrate that granular mortality rates inferred from the FamiLinx genealogical dataset are, in general, not representative. We use a statistical model to estimate the adjustment factors to the FamiLinx-derived mortality rates by comparing to data from the Human Mortality Database. Age-specific mortality rates inferred from the genealogy are too low for younger ages, but are more representative in mid- and late-adult ages, though there is considerable variation over time and geography. The rich range of patterns is able to be captured by the model's flexible specification of the adjustment factor. Extrapolation to populations without reference data is enabled by enforcing regular shapes on the mortality curve, which allow us to indirectly identify novel patterns in the adjustment factor even for out-of-sample countries.

The combination of a set of key features of the model allow us to separate the mortality rate μ from the adjustment factor ψ. The choice of mortality model is important. While there have been other models proposed for regularizing mortality estimates, typically for small area mortality estimation, we adapt the approach from Alexander *et al.* (2017) because it is cast in a Bayesian framework, and critically assumes that most variation occurs along the principal axes, which restricts the possible mortality shapes. Under a TOPALS-like approach such as that from Schmertmann and Gonzaga (2018), the shape is much more flexible (as long as variation from the standard mortality curve is smooth), which in turn leads to poor identification of ψ.

Unlike the Normally-distributed error terms in Alexander *et al.* (2017), we accommodate deviations from the shape using a sparse set of ε terms with regularized horseshoe priors. The sparsity here is needed so that ε does not compete with μ and ψ to explain systematic variation, but since those components of the model are smoothed, the only way for the model to explain isolated shocks is through ε. It is interesting to note that this setup was only feasible to compute because of recent but unrelated methodological developments, namely the regularized horseshoe prior of Piironen and Vehtari (2017) and the tensor product spline parameterization of Wood *et al.* (2013) both eased computation.

5.1 Bias in the FamiLinx dataset

Our findings of substantial bias in the FamiLinx-inferred mortality rates could be seen as contrary to claims of representativeness made by Kaplanis *et al.* (2018). We offer several possible explanations for this conflict. First, the rules for data inclusion differ. In their paper, the authors include only profiles which contained exact birth and death dates to avoid age heaping, which also may have resulted in an overall higher quality sample in some respects, but may be biased with respect to socioeconomic composition (Stelter and Alburez-Gutierrez, 2022). In our case, to obtain counts large enough to yield stable rates, we relax this requirement to use only birth (or baptism) and death (or burial) year, and work in 5-year periods to mitigate age-heaping effects. Second, geographic information was treated differently. Kaplanis *et al.* (2018) use a geoparsing algorithm to assign coordinates to free text, whereas we use a simpler string matching approach for country assignment. Third, we present mortality as rates and deduce life expectancies, whereas the original paper considers lifespans and distributions of age at death. Lastly, here we have presented specific national mortality rates, as opposed to aggregated statistics over several countries. The sensitivity of the substantive conclusion to the set of analysis choices calls for careful consideration of the research target or estimand, relevant subset of the data, and assumptions and simplifications introduced in future analyses.

Differences in the mortality rate bias between countries and periods suggest that they are affected differentially by mechanisms in the data-generating process. As such, it is difficult to say with generality how to account for deficiencies in these data for a particular population and estimand. Given the distinctive context of the United States for example, it is unsurprising that the adjustment factor is dissimilar to those seen in other countries, but prior to this study, there was no way to know the particular extent and age-temporal structure of the bias without having reference data. Our strategy here works because of a rigid mortality model that forces the discrepancies to be explained by ψ, which is sensible for mortality but will apply to varying extents for estimation of other demographic quantities depending on the strength of regular patterns that can be exploited.

However, the results we present here may provide clues for investigating biases in other demographic quantities of interest. One can imagine, for example, that the extent of infant mortality underreporting is perhaps suggestive of the extent of fertility underreporting. Another hypothesis is that

the distinctive change over time in Great Britain could be a consequence of the availability of information for migrants and non-migrants. If disproportionately many of the database contributors are descendants of migrants from Great Britain, then they may record their direct ancestors at a higher rate than those who did not migrate. Under our calculation of deaths and exposure, they would contribute to the exposure for the British mortality rate, but not to the count of deaths. We hope that the present contribution can serve as a useful starting point for more detailed assessments of plausible deficiencies in this rich dataset.

5.2 Limitations of the method and possible extensions

Further attention might be paid to shocks in mortality due to significant historical events. The smoothing of the adjustment factor means that acute jumps in the FamiLinx mortality rate are reflected almost wholly in the overdispersion of the negative binomial, or the ε terms in the mortality rate, but there could be anomalous changes to access to knowledge for that period. During high mortality events such as war, we might expect that many soldiers die before having children and may therefore be subject to missingness due to lack of descendants, but for a crowdsourced genealogy it could be that users of the genealogy may be especially motivated to record deaths due to significant historical events, resulting in relatively high reporting of mortality in that period. A possible avenue to detect these reporting anomalies it could be useful to look at the fertility of the periods during which the affected cohort was born.

Concerning our modelling approach, we are cautious regarding the generalizability of the model to countries very different from those with reference data, which is a consideration shared by nearly all structured mortality estimation methods. If populations persistently do not adhere to the patterns allowed by the mortality model, it is difficult to capture the true shape of mortality, as in certain periods of the validation results in Sweden. It may also be possible, as in the case of Great Britain, for there to be a way for the model parameters to explain the FamiLinx data that does not reflect the true patterns.

Therefore, for future studies targeting specific historical populations, we believe estimates can be improved by integrating evidence from other sources and our model has several places where this can be done. Given more specific historical information about a population, one could make more

informed choices of reference populations from which to draw the principal axes for mortality, which would encourage the model to follow hypothesized mortality patterns. Incorporating other observations, if available, can be also done by relating the data to μ through a likelihood. One could also choose to impose more informative priors or pool information between different populations in the mortality part of the model, if it is believed that they followed similar trajectories or patterns of mortality.

It is worth noting that while our method is designed to adjust for the biases that arise from the genealogy, we have not informed our model with any characteristics of the genealogy. As our understanding of these data improves, researchers may be able to better hypothesize about how to detect deficiencies and form more informative priors on the adjustment factor.

5.3 Interpretation of estimates for the United States

We estimate that life expectancy in the United States was higher in the early 19th century and lower in the late 19th century than previously thought by Hacker (2010), which has been, to our knowledge, the most thorough attempt thus far to estimate mortality in this period. We also find that the life expectancy decline in the first half of the century was more dramatic in the male population, and the late-century life expectancy increase was less pronounced in both the male and female population. Interestingly, the estimates of Hacker (2010) is also based on adjusting data from genealogies, but there are important conceptual differences between our approaches.

In general, our method is driven more by data, relaxing several assumptions underlying Hacker's method. First, our choice of mortality model is far more flexible compared to the model life table approach, which posits very rigid relationships between age and mortality. Instead of assuming a particular age structure, our model can learn and deviate from known shapes of mortality based on patterns in the data. We are also able to estimate male and female mortality separately. Due to data limitations for the female population, Hacker assumes a fixed difference between male and female life expectancy in order to then obtain estimates of female mortality, which does not allow for the identification of separate trends for the male and female populations.

Given the distinctive context of the United States however, there are opportunities to refine our

results further. While past efforts (including Hacker (2010)) have focused on the white population[3], the racial composition of the population in FamiLinx is less clear because of the lack of identifying information in the dataset. More research is needed to better understand the extent to which this may account for the observed differences, and how to improve estimates for non-white populations. Furthermore, Hacker (2010) makes a case for understanding U.S. mortality through urban and rural populations' life tables, though he also points out that the available U.S. life tables from 1900-1902 may not reflect important historical trends seen in 19th-century Europe. Even if the truth were somewhere in between, drawing on these sources for the principal axes could allow the model to more parsimoniously explain mortality trends.

5.4 Conclusion

We conclude with a summary of implications for future research and potential directions. From a statistical perspective, it would be interesting to consider how to adapt the modelling strategy of combining smoothing techniques and sparsity-inducing techniques for other contexts. The recent methodological advancements both these fronts that we have applied here have made such problems more computationally tractable, and we imagine that there are many other applications where this could apply.

From a demography perspective, the model we have described here could be easily adapted to fit into a larger estimation framework for historical mortality that incorporates other data sources. This can help reduce uncertainty and address the limitations from any singular imperfect data source. We also highlight that the utility of regularizing mortality estimates in our model is not just for the sake of sensible mortality estimates, but also to disentangle the various parts of the data-generating process in the FamiLinx dataset.

Our approach could also be extended to other quantities, such as fertility and migration, or perhaps even to estimate them jointly with mortality, although regularization of the estimands will likely require more consideration. In lieu of equivalent historical data to draw upon, it may be interesting to consider how relationships from formal demography could be used to provide structure for the

[3]Estimates for the black population are sometimes presented (e.g. by Haines (1994)) but the choice of model life table is difficult to conclusively validate.

Table A1: Modifications to set of expressions used to match country names.

Country	Modification
United States of America	We allowed the following expressions to be matched:'UNITED STATES'; 'UNITED STATES OF AMERICA'; ', AMERICA'; 'REPUBLIC OF AMERICA'; 'NEW ENGLAND'; 'UNITED COLONIES OF AMERICA'; 'MASSACHUSETTS'; 'CONNECTICUT'; 'ILLINOIS'; 'NEW YORK'; 'NORTH CAROLINA'; 'SOUTH CAROLINA'; 'NEW HAMPSHIRE'; 'KENTUCKY'; 'RHODE ISLAND'; 'COLONIAL AMERICA'; 'PROVINCE OF VIRGINIA'; 'TEXAS'; 'CALIFORNIA'
France	We removed strings that included 'FRANCE', but added: '\w*(?<!NEW)FRANCE'; '\w*(?<!NOUVELLE[-])FRANCE'; 'R(É|E)P.*FRANÇAISE'; 'FRENCH.?REPUBLIC'
Netherlands	We removed strings that matched '^OLAND$|^HOLLAND*', but added '\w*(?<!NEW)HOLLAND')
Great Britain	We removed 'ENGLAND', and added:'BRITAIN'; '\w*(?<!NEW)SCOTLAND'; ' UK'; ' GBR'; '\w*(?<!NEW)ENGLAND'; '\w*(?<!NEW SOUTH)WALES'
Sweden	We added 'SUÃ¨DE'
Finland	We added 'SOOMLANE'

estimands.

Finally, as we have demonstrated in this study, users of online genealogies for research should be aware of the potentially very serious selection issues in these data. The heterogeneity and complexity of these effects across populations present significant challenges, but methodological investment here can have large returns for studying big demographic questions empirically. Our work constitutes just one step towards making online genealogies more useful for the demographic community.

A Appendix: Country extraction

For some countries, we make modifications to the set of names and expressions in the `countrycode` package to prevent false matches or accommodate certain data entry mistakes. These are specified in Table A1 below.

B Appendix: Migration composition of deaths and exposure-to-death

This appendix contains plots that show the composition of information contributing to the death counts and exposure-to-death. For each (country-age-sex specific) population, the death count consists of those where the location of death was specified, and those where the location of death was imputed from the location of the most recent vital event, whereas the exposure-to-death is composed of segments of individuals lives with possibly missing or different location information at the endpoints. See Section 2.2 and Figure 1 for details.

B.1 Migration composition of death counts

Figure A1: Composition of death counts in Belgium.

Figure A2: Composition of death counts in Switzerland.

Figure A3: Composition of death counts in Denmark.

Figure A4: Composition of death counts in Finland.

Figure A5: Composition of death counts in France.

Figure A6: Composition of death counts in Great Britain.

Figure A7: Composition of death counts in Netherlands.

Figure A8: Composition of death counts in Norway.

Figure A9: Composition of death counts in Sweden.

Figure A10: Composition of death counts in United States.

B.2 Migration composition of exposure-to-death

Figure A11: Composition of exposure-to-death in Belgium.

Figure A12: Composition of exposure-to-death in Switzerland.

Figure A13: Composition of exposure-to-death in Denmark.

Figure A14: Composition of exposure-to-death in Finland.

Figure A15: Composition of exposure-to-death in France.

Figure A16: Composition of exposure-to-death in Great Britain.

Figure A17: Composition of exposure-to-death in Netherlands.

Figure A18: Composition of exposure-to-death in Norway.

Figure A19: Composition of exposure-to-death in Sweden.

Figure A20: Composition of exposure-to-death in United States.

C Appendix: Validation

(a) Belgium female population.

(b) Belgium male population.

Figure A21: Leave-data-out validation exercise results for Belgium.

(a) Switzerland female population.

(b) Switzerland male population.

Figure A22: Leave-data-out validation exercise results for Switzerland.

(a) Denmark female population.

(b) Denmark male population.

Figure A23: Leave-data-out validation exercise results for Denmark.

(a) Finland female population.

(b) Finland male population.

Figure A24: Leave-data-out validation exercise results for Finland.

(a) France female population.

(b) France male population.

Figure A25: Leave-data-out validation exercise results for France.

(a) Great Britain female population.

(b) Great Britain male population.

Figure A26: Leave-data-out validation exercise results for Great Britain.

(a) Netherlands female population.

(b) Netherlands male population.

Figure A27: Leave-data-out validation exercise results for Netherlands.

(a) Norway female population.

(b) Norway male population.

Figure A28: Leave-data-out validation exercise results for Norway.

(a) Sweden female population.

(b) Sweden male population.

Figure A29: Leave-data-out validation exercise results for Sweden.

D Appendix: Regularized horseshoe hyperprior sensitivity

In our mortality model described in Section 3.2.2, the sparsity of the deviations ε away from the structured part of the mortality model is controlled by the regularized horseshoe prior

$$\varepsilon_i \sim N(0, \tau^2 \tilde{\lambda}_i^2), \tau \sim \text{Cauchy}(0, \tau_0)$$

where we choose τ_0 to be 0.01. We test the sensitivity of our analysis to the choice of τ_0 by fitting our model when τ_0 is set to the alternative values of 0.001, 0.1, and 1. Table A2 shows the resulting estimates of τ and Figure A30 shows the distribution of ε_i under these settings.

Table A2: Estimates of τ under different settings of τ_0.

τ_0	Posterior mean of τ	Posterior s.d. of τ
0.001	0.03032	0.001859
0.010	0.03082	0.001855
0.100	0.03057	0.001996
1.000	0.03044	0.001943

The results suggest that our analysis is not very sensitive to the choice of τ_0. Estimates of both τ and ε_i are largely unchanged.

E Appendix: Details of adjustment factor spline prior

Under the construction of Wood *et al.* (2013), the second term relating to Z actually consists of three partitions $Z = [Z_1 : Z_2 : Z_3]$, each to be treated as a block of i.i.d. random effects (see Section

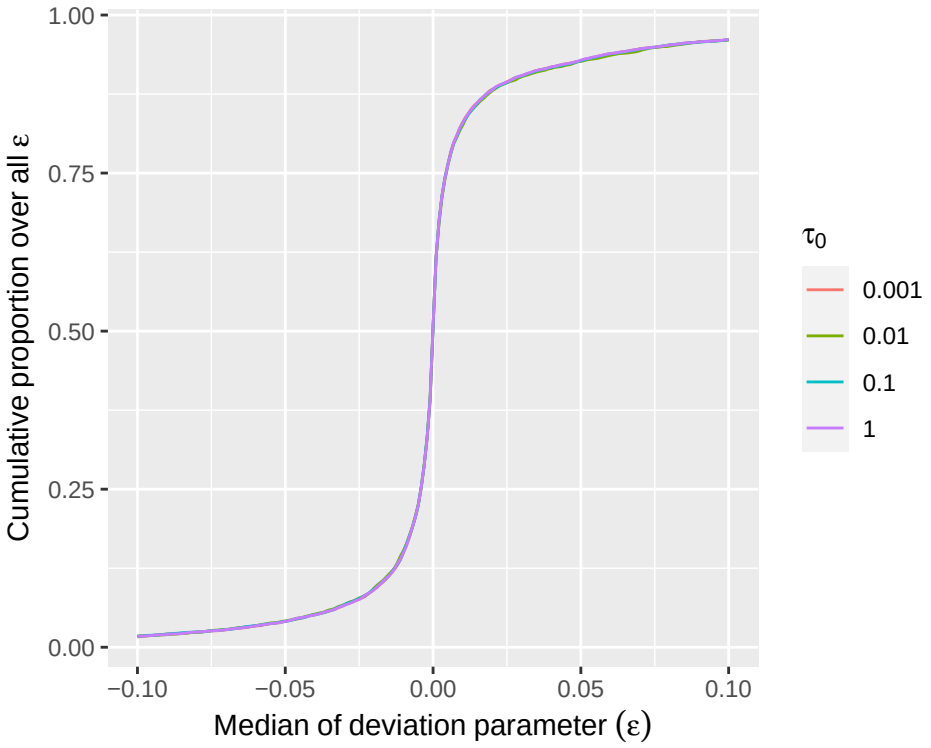


Figure A30: Distribution of ε_i under different settings of τ_0. Each line shows the cumulative distribution of the set of ε_i under a different τ_0.

3 and Appendix 2 of their paper for details). We therefore write Equation 1 more explicitly as

$$\vec{f}_{c,g} = X\vec{\beta}_{c,g} + Z_1\vec{\delta}^{(1)}_{c,g} + Z_2\vec{\delta}^{(2)}_{c,g} + Z_3\vec{\delta}^{(3)}_{c,g}. \tag{2}$$

and for each of the components $j \in \{1, 2, 3\}$, $\vec{\delta}^{(j)}_{c,g}$ is defined as

$$\vec{\delta}^{(j)}_{c,g} = \vec{\delta}^{(j)}_{g} + \vec{\zeta}^{(j)}_{c,g}.$$

The set of $\vec{\delta}^{(j)}_{g}$ determines a global nonlinear trend for each gender, whereas the set of $\vec{\zeta}^{(j)}_{c,g}$ represents country-specific variation from those global trends for each gender. The former are subject to

$$\vec{\delta}^{(j)}_{g} \sim MVN(0, \sigma^2_{\delta^{(j)}}\mathbf{I}), \tag{3}$$

while the latter are subject to

$$\vec{\zeta}^{(j)}_{c,g} \sim MVN(0, \sigma^2_{\zeta,j,g}\mathbf{I}). \tag{4}$$

The variance parameters in Equation 3 and Equation 4 control the amount of variation allowed from the linear trends, or the 'wiggliness' of the adjustment factor. Each of the standard deviation parameters is given an $N^+(0, 1^2)$ prior:

$$\sigma_{\delta^{(j)}} \sim N^+(0, 1^2) \text{ for } j \in \{1, 2, 3\},$$

and

$$\sigma_{\zeta,j,g} \sim N^+(0, 1^2) \text{ for } j \in \{1, 2, 3\}.$$

References

Alexander, M., Zagheni, E. and Barbieri, M. (2017) A flexible bayesian model for estimating subnational mortality. *Demography*, **54**, 2025–2041. Duke University Press.

Arel-Bundock, V., Enevoldsen, N. and Yetman, C. (2018) Countrycode: An r package to convert country names and country codes. *Journal of Open Source Software*, **3**, 848. Available at: https://doi.org/10.21105/joss.00848.

Blanc, G. (2020) Modernization Before Industrialization: Cultural Roots of the Demographic Transition in France. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.3702670.

Calderón-Bernal, L. P., Alburez-Gutierrez, D. and Zagheni, E. (2023) *Analyzing biases in genealogies using demographic microsimulation*. Max Planck Institute for Demographic Research, Rostock, Germany.

Clark, S. J. (2019) A general age-specific mortality model with an example indexed by child mortality or both child and adult mortality. *Demography*, **56**, 1131–1159. Duke University Press.

Colasurdo, A. and Omenti, R. (2024) Using online genealogical data for demographic research: An empirical examination of the FamiLinx database. SocArXiv. DOI: 10.31235/osf.io/62yxm.

Dharamshi, A., Alexander, M., Winant, C., et al. (2023) Jointly estimating subnational mortality for multiple populations. *arXiv preprint arXiv:2310.03113*.

Gabry, J., Češnovar, R. and Johnson, A. (2023) *Cmdstanr: R Interface to 'CmdStan'*.

Gonzaga, M. R. and Schmertmann, C. P. (2016) Estimating age-and sex-specific mortality rates for small areas with TOPALS regression: An application to brazil in 2010. *Revista Brasileira de Estudos de População*, **33**, 629–652. SciELO Brasil.

Hacker, J. D. (2010) Decennial life tables for the white population of the united states, 1790–1900. *Historical methods*, **43**, 45–79. Taylor & Francis.

Haines, M. R. (1979) The use of model life tables to estimate mortality for the United States in the late nineteenth century. *Demography*, **16**, 289–312. Duke University Press.

Haines, M. R. (1994) *Estimated life tables for the united states, 1850-1900*. Historical working paper series 59, Working Paper. National Bureau of Economic Research. DOI: 10.3386/h0059.

Haines, M. R. (2001) The urban mortality transition in the United States, 1800-1940. *Annales de démographie historique*, **n^o 101**, 33–64. CAIRN. DOI: 10.3917/adh.101.0033.

Henry, L. (1956) Anciennes familles genevoises. Etude démographique: XVIme - XXme siècle. *Population (French Edition)*, **11**, 334. DOI: 10.2307/1524668.

Holden, L. and Boudko, S. (2018) The Norwegian Historic Population Register and Migration. *Journal of Migration History*, **4**, 249–263. DOI: 10.1163/23519924-00402002.

Hollingsworth, T. H. (1976) Genealogy and historical demography. *Annales de démographie historique*, **1976**, 167–170. DOI: 10.3406/adh.1976.1310.

Human Fertility Database (2024) Human Fertility Database. Max Planck Institute for Demographic Research (Germany); Vienna Institute of Demography (Austria).

Human Mortality Database (2024) Human Mortality Database. Max Planck Institute for Demographic Research (Germany); University of California, Berkeley (USA); French Institute for Demographic Studies (France).

Kaplanis, J., Gordon, A., Shor, T., et al. (2018) Quantitative analysis of population-scale family trees with millions of relatives. *Science*, **360**, 171–175. DOI: 10.1126/science.aam9309.

Pedersen, E. J., Miller, D. L., Simpson, G. L., et al. (2019) Hierarchical generalized additive models in ecology: An introduction with mgcv. *PeerJ*, **7**, e6876. PeerJ Inc.

Piironen, J. and Vehtari, A. (2017) Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, **11**, 5018–5051.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Schmertmann, C. P. and Gonzaga, M. R. (2018) Bayesian estimation of age-specific mortality and life expectancy for small areas with defective vital records. *Demography*, **55**, 1363–1388. Duke University Press.

Stelter, R. and Alburez-Gutierrez, D. (2022) Representativeness is crucial for inferring demographic processes from online genealogies: Evidence from lifespan dynamics. *Proceedings of the National Academy of Sciences*, **119**, e2120455119. DOI: 10.1073/pnas.2120455119.

Wilmoth, J., Zureick, S., Canudas-Romo, V., et al. (2012) A flexible two-dimensional mortality model for use in indirect estimation. *Population studies*, **66**, 1–28. Taylor & Francis.

Wood, S. N. (2004) Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, **99**, 673–686.

Wood, S. N. (2017) *Generalized Additive Models: An Introduction with r*. 2nd ed. Chapman; Hall/CRC.

Wood, S. N., Scheipl, F. and Faraway, J. J. (2013) Straightforward intermediate rank tensor product smoothing in mixed models. *Statistics and Computing*, **23**, 341–360. Springer.

Wrigley, E. A. and Schofield, R. S. (1983) English Population History from Family Reconstitution:

Summary Results 1600-1799. *Population Studies*, **37**, 157. DOI: 10.2307/2173980.

Zhao, Z. (2001) Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies*, **55**, 181–193. DOI: 10.1080/00324720127690.