MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Konrad-Zuse-Strasse 1 · D-18057 Rostock · Germany · Tel +49 (0) 3 81 20 81 - 0 · Fax +49 (0) 3 81 20 81 - 202 · www.demogr.mpg.de

MPIDR Working Paper WP 2022-005 I January 2022
https://doi.org/10.4054/MPIDR-WP-2022-005

Identifying and correcting bias in big crowd-sourced online genealogies

Michael Chong
Diego Alburez-Gutierrez I alburezgutierrez@demogr.mpg.de
Emanuele Del Fava I delfava@demogr.mpg.de
Monica Alexander
Emilio Zagheni I office-zagheni@demogr.mpg.de

Identifying and correcting bias in big crowd-sourced online genealogies

Michael Chong[1], Diego Alburez-Gutierrez[2], Emanuele Del Fava[2], Monica Alexander[1,3], and Emilio Zagheni[2]

[1]Department of Statistical Sciences, University of Toronto
[2]Laboratory of Digital and Computational Demography, Max Planck Institute for Demographic Research
[3]Department of Sociology, University of Toronto

Abstract

Human societies have long valued genealogies as repositories of historical information. However, historical research has shown that genealogies are flawed and should not be taken at face-value. In recent years, online communities have produced big genealogies connecting all continents over multiple centuries. We present the first attempt to characterize and account for systematic biases in demographic rates inferred from online genealogy profiles. We construct a Bayesian model to compare life event data in a crowd-sourced genealogical dataset with data from the Human Mortality Database (HMD) in four European countries. We find that mortality in the genealogy data is under-reported, and propose a method to adjust mortality rates by calibrating against high quality data. To evaluate the method, we estimate out-of-sample adjusted mortality rates for 1835-1900 Finland, and find that they are much closer to the HMD over the period where HMD estimates are available (1878 onward). Finally, we obtain adjusted mortality rates for the United States 1835-1900 for which, to our knowledge, no high quality data exist. We expect this to be the first of many studies to harvest online genealogical data to improve our understanding of historical human dynamics.

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 &

Schofield, 1983) and ascendant genealogies, where individuals reconstruct their ancestry retrospectively, but increased data availability has allowed demographers to use large genealogical datasets to study demographic change directly (Zhao, 2001; Holden & Boudko, 2018). In recent years, the spread of 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 over multiple countries and centuries (Kaplanis et al., 2018). These data, which have been curated by a team of computational scientists to remove duplicates and other inconsistencies, have many potential applications for demographic analysis.

The Familinx dataset contains individual-level information on the place and date of birth and death of 86 million individuals. It has an unparalleled chronological and geographic coverage, encompassing most of the Western World over the past five centuries. However, its main strength is the fact that it records kinship ties between parents and children, which can be used to reconstruct complete genealogies. Crucially, unlike register-based genealogies, kin ties are not restricted by national borders, making Familinx a truly transnational data source. The data could shed unique new insights into long-term demographic trends, such as the evolution of longevity and lifespan inequality (Oeppen, 2002; van Raalte, Sasson, & Martikainen, 2018). Long-term kinship networks allow for the study of demographic change from the point of view of kin (Murphy, 2004). These include questions of the relationship between demographic and kinship transitions (Murphy, 2011; Verdery, 2015) and the hereditability of demographic behaviour (Kolk, 2014). These data can also contribute to the growing interest in understanding both the prevalence and consequences of the experience of kin death (Alburez-Gutierrez, Kolk, & Zagheni, 2021).

It may seem surprising that such a rich data source has not been used extensively for demographic research. Despite their great potential, online genealogies include many biases that restrict their usability (Hollingsworth, 1976). First, the production of genealogies is contingent on historical and social forces and, as a result, family histories tend to under-represent women, early deaths, and marriages without children, to name only a few (Zhao, 2001). In addition to this, ascendant genealogies like the one we consider in this paper might also be biased by design. The Familinx data, for example, is the product of thousands of genealogists attempting to retrospectively reconstruct their own family histories. As a result, any given individual's inclusion in the Familinx data is conditional on having a living descendant with an interest in genealogies and may be affected by processes of 'selective remembering', whereby some ancestors are deemed more worthy of being included in a family tree than others. This suggests the presence of selection effects such as survivorship bias and lineage

extinction (Zhao, 2001).

In spite of these serious concerns regarding data quality, existing studies using online genealogies have taken the data at face-value, without attempting to account for any source of bias (Fire & Elovici, 2013). Indeed, studies have often assumed that the data is representative of the broad population (Kaplanis et al., 2018; Blanc, 2020). As we will argue in this paper, this may not be true for particular sub-populations and time periods. Other areas in the emerging field of digital demography have seen the introduction of methods for correcting digital bias (Alexander, Polimis, & Zagheni, 2020), but similar work has not been developed yet for online genealogies. Researchers at the intersection of demography, statistics, and computer science are in a privileged position to address this problem.

In this paper, we quantify the bias in mortality rates derived from online genealogies and propose a statistical methodology to account for it. Using historical data from the Human Mortality Database as a benchmark, we fit a Bayesian model to estimate the structure and degree of misrepresentation of mortality rates in the Familinx data. This is captured in a set of 'weights' that can then be used to adjust mortality rates estimated from online genealogies. The development of such bias-correction techniques is a first step towards unlocking the potential of online genealogies to conduct quality demographic research on a wide range of sociological and demographic issues including, but not restricted to, mortality, fertility, migration, and intergenerational processes.

The rest of this paper is structured as follows. First, we introduce the two datasets used in the analysis, i.e., the Familinx online genealogies and the Human Mortality Database. Following this, we clarify how we processed the genealogical data to extract age-specific mortality rates. We then introduce our modelling approach and explain how it was applied to model and correct the systematic bias in the online genealogies. We exemplify our results for a selection of countries, highlighting the potential of our correction mechanism to improve historical estimates derived from online genealogies. After evaluating the strengths and shortcomings of our model, we 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

We consider online genealogies that come from Familinx, a 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 including individuals born over the last 400 years on all continents, although data quality varies greatly by time period and region. The site users who contributed the data mainly come from countries in the Global North, and this is reflected in the location of the recorded vital events, 55% of which are located in Europe and 30% in North America. We computed period sex and age-specific mortality rates from the raw Familinx data following standard demographic procedures. 'Gold-standard' period historical demographic rates come from the Human Mortality Database (HMD), a widely used and high-quality repository of harmonized demographic data.

3 Methods

3.1 Extraction

Obtaining country and period-specific death counts and exposure from the Familinx dataset is challenging for a number of reasons. First, the time and location information on an individual's vital events is often incomplete. Approximately 60% of profiles are missing their year of birth, and among profiles with years of birth before 1900, approximately 42% are missing year of death. Second, geographic information is often free-text, and therefore contains errors, historical (defunct) names, and names given in the country's (non-English) language. In their original paper, Kaplanis et al. (2018) used a geocoding service to assign countries to profiles. Since this step may be cost-prohibitive to some researchers, we pursue a simpler string-matching approach to country assignment (we provide details later on and in the appendix). Third, there is no information on the place of residence of individuals over the course of their life, therefore requiring assumptions about migration to obtain country-specific exposure-to-death.

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 im-

plausible implied 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 6 countries that we consider (Denmark, Finland, France, Norway, Sweden, United States) by searching for a set of location names in location-relevant data fields. Details on the names used for matching to these countries is given in Appendix A.1. For simplicity, we assume that individuals do not migrate between countries. An individual is assigned to their country of death or their country of birth in that order of preference. To illustrate, if an individual was born in Denmark and died in Sweden, they would count only toward Sweden's death and exposure-to-death. If an individual was born in Denmark, but died in Canada, they would not be counted, since the death location is preferred, and Canada is not under consideration. If an individual was born in Denmark but there was no information on location of death, then they would count toward Denmark's death and exposure-to-death.

The above scheme was chosen as a trade-off between parsimonious assumptions and capturing a greater number of profiles, the latter of which is especially important where counts are small and noisy. Figure 1 shows how the inferred number of deaths changes under different modifications of the current scheme (labelled as Scheme 3). We explored more conservative assumptions, such as only using records where birth and death place agree (Scheme 1) or only using death place (Scheme 2), and less conservative imputations such as using the birthplace of children if they all agree (Scheme 4), using the birthplace of the most recent child (Scheme 5) and using the parents' birth and death place if they all agree (Scheme 6). Using partial information in the profile greatly increases the number of usable profiles (Schemes 1-3), but further using relatives' location information (Schemes 4-6) does not.

3.2 Statistical model

3.2.1 Model overview

Our statistical model quantifies the discrepancy, or bias, between the mortality rates derived from the genealogy and those from the HMD, which are assumed to be of high quality. An estimate of the bias comes from country-periods where there is high-quality data available (hereafter, "training countries"). The difference between the rates is captured by an adjustment factor, denoted ψ, which represents the ratio of the genealogy-derived mortality rate

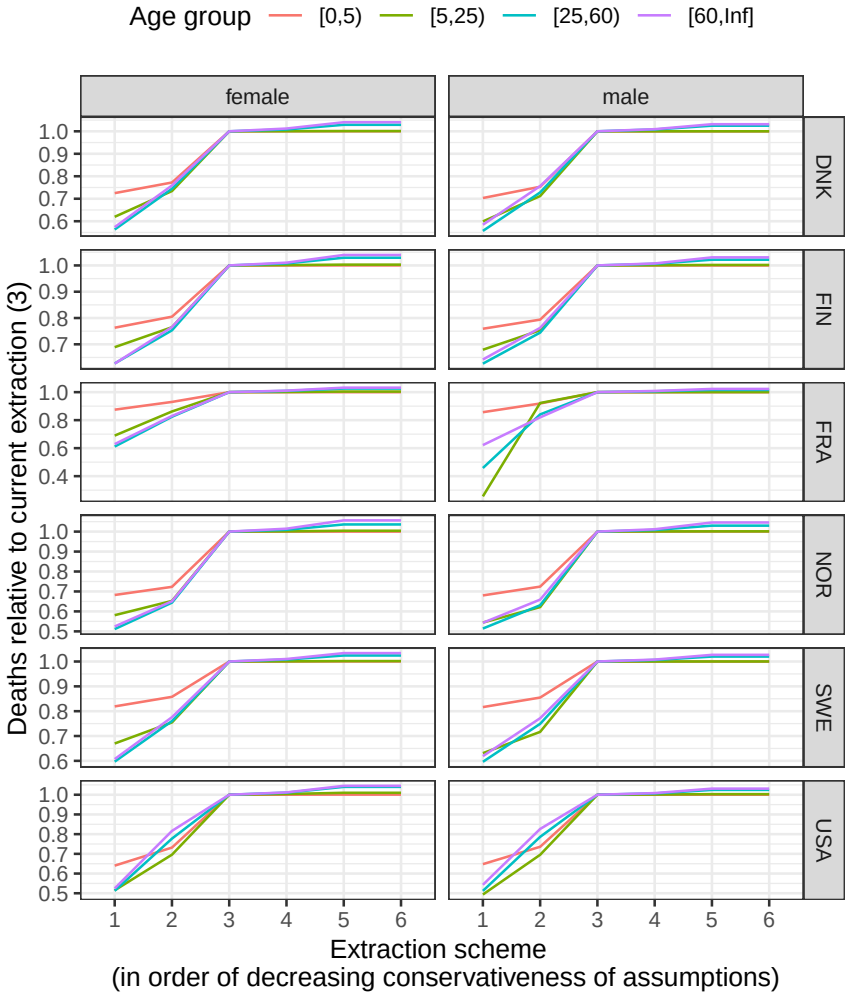


Figure 1: Relative number of deaths captured under different imputation assumptions. We use Scheme 3 for the rest of the analysis, in which an individual is assigned to (in order of preference) their death place or birth place. Extraction schemes 1 and 2 are more conservative with respect to country assignment, while 4-6 are less conservative.

to the HMD rate. The adjustment factor is allowed to vary by age, cohort, and a set of country/age/period-specific indicators of data quality in that subset of the data.

The estimate of bias is then used to correct the genealogy-derived mortality rates in country-periods where there is no high-quality data available (hereafter "test countries"). Since the adjustment factor ψ represents the ratio of mortality rates, a corrected rate is yielded by dividing the genealogy-derived rate by the adjustment factor.

3.2.2 Deaths process

For training country c, sex s, age group x, and period t, let d_{csxt} and p_{csxt} denote the corresponding number of deaths and exposure respectively in the Familinx data. Let $m_{csxt} = d_{csxt}/p_{csxt}$ denote a naive genealogy-derived mortality rate, and let M_{csxt} denote the corresponding HMD mortality rate.

Deaths are assumed to be Poisson-distributed with rate $p_{csxt} \cdot M_{csxt} \cdot \psi_{csxt}$.

$$d_{csxt} \sim \text{Poisson}(p_{csxt} \cdot M_{csxt} \cdot \psi_{csxt}).$$

The parameter ψ denotes an adjustment factor which captures the ratio of the Familinx mortality rate to the HMD rate. That is, ψ is interpreted as m/M.

3.2.3 Structure on adjustment factor

The adjustment factor ψ_{csxt} is then modelled according to the following equation:

$$\log(\psi_{csxt}) = \alpha_{sx} + \gamma_{sb(xt)} + X_{csxt}\vec{\beta}_{sx}, \tag{1}$$

where α_{sx} represents an age group effect for age x, $\gamma_{sb(xt)}$ denotes a cohort effect for birth year $b(xt)$ of individuals aged x at time t, and X_{csxt} denotes a set of covariates with age-varying coefficients $\vec{\beta}_{sx} = (\beta_{sx1}, \ldots, \beta_{sxK})$. Note that the model is fit separately for the male and female populations, meaning all parameters are sex-specific. We also investigated using a period effect instead of a cohort effect, but found that the model with a cohort effect performed slightly better, and including all three (age, period, cohort) would lead to identifiability issues.

3.2.4 Covariates

We include a set of three covariates that summarize certain characteristics of the data in order to capture variation in data quality in the genealogy, which may then inform variation in the mortality rate bias. First, we use the logged ratio of the male population to the female population, calculated by counting individuals assigned to that country who have birth year, death year, and sex recorded. Second, we use the logit-transformed proportion of records with no death date recorded. The numerator of this proportion is taken to be the number of individuals of the relevant sex and country who have a birth year recorded, no death year recorded, and is younger than 90 years old. The denominator is calculated as the sum of this term and the number of records with a death recorded. Finally, we also include the logit-transformed proportion of records without a known parent, among individuals of that sex and country with a birth and death date recorded.

3.2.5 Calculation of adjustment factors for test countries

Adjustment factors $\hat{\psi}_{c^*sxt}$ for a test country c^* are estimated according to the following equation:

$$\log \hat{\psi}_{c^*sxt} = \alpha_{sx} + \gamma_{sb(xt)} + X_{c^*sxt}\vec{\beta}_{sx} + \varepsilon_{c^*sxt}, \tag{2}$$

where the parameters α, γ, and $\vec{\beta}$ are the same as those above in Equation 1, and X_{c^*sxt} are the covariates as calculated for the test country. An additional error term ε_{c^*sxt} is used to allow extra flexibility in the resulting mortality curve to accommodate the smoothing discussed below in Section 3.2.6.

Recalling that the adjustment factor ψ captures the ratio of the Familinx rate m to the HMD rate M, we obtain an adjusted mortality estimate $\hat{M}_{c^*sxt}$ via the relation

$$\hat{M}_{c^*sxt} = \frac{m_{c^*sxt}}{\psi_{c^*sxt}}.$$

3.2.6 Smoothing of resulting mortality curve

If left as-is, the resulting mortality rates M_{c^*sxt} may have some unrealistic characteristics. For instance, the mortality rates at the oldest age group are unstable due to low counts and can be unusually low. Noisiness in the raw Familinx mortality curve may also propagate to the adjusted curve, resulting in overly jagged estimates over age. To mitigate these issues,

the adjusted mortality curve is made subject to a prior which incorporates information about the "expected shape" of the curve. The adjusted mortality curve is modelled as

$$\log \hat{M}_{c^*sxt} \sim \text{Normal}(\mu_{sxt}, \sigma_\mu^2), \tag{3}$$

where μ_{sxt} are defined as entries in the vector

$$\begin{bmatrix} \mu_{s1t} \\ \mu_{s2t} \\ \vdots \\ \mu_{sXt} \end{bmatrix} = \eta_{s1t} \cdot \vec{v}_1 + \eta_{s2t} \cdot \vec{v}_2.$$

Here η_{s1t} and η_{s2t} are parameters to be estimated, and $\vec{v}_1$ and $\vec{v}_2$ are the first two right singular vectors from the singular value decomposition (SVD) of the training countries' log HMD mortality rates. More explicitly, if UDV^T is the SVD of the matrix

$$\begin{bmatrix} \log M_{1s11} & \log M_{1s21} & \cdots & \log M_{1sA1} \\ \log M_{1s12} & \log M_{1s22} & \cdots & \log M_{1sA2} \\ \vdots & \vdots & \vdots & \vdots \\ \log M_{1s1T} & \log M_{1s2T} & \cdots & \log M_{1sAT} \\ \log M_{2s11} & \log M_{2s21} & \cdots & \log M_{2sA1} \\ \vdots & \vdots & \vdots & \vdots \\ \log M_{Cs1T} & \log M_{Cs2T} & \cdots & \log M_{CsAT} \end{bmatrix},$$

then $\vec{v}_1$ and $\vec{v}_2$ are respectively the first and second columns of V. The estimated parameters η_{s1t} and η_{s2t} follow the distribution

$$\eta_{s1t} \sim \text{Normal}(\bar{\eta}_{s1}, 10^2)$$
$$\eta_{s2t} \sim \text{Normal}(\bar{\eta}_{s2}, 10^2)$$

where $\bar{\eta}_{s1}$ and $\bar{\eta}_{s2}$ are the means of the entries of first and second columns of the matrix UD respectively.

The right singular vectors $\vec{v}_1$ and $\vec{v}_2$ are given in Appendix A.6. We note that the first vector, which captures a baseline shape of the curves, already represents the vast majority of the structure found in the log mortality rates. This is evidenced by the relative magnitude of the mean "coefficient" for this vector, $\bar{\eta}_{s1} \approx -17.1$, which is much larger in magnitude than any of the subsequent ones ($\bar{\eta}_{s2} \approx 0.005$).

3.2.7 Priors

The age effect, cohort effect, and coefficients are assigned independent weakly informative priors, i.e.,

$$\alpha_{sx} \sim \text{Normal}(0, 10^2)$$
$$\gamma_{sb(xt)} \sim \text{Normal}(0, 10^2)$$
$$\beta_{sxk} \sim \text{Normal}(0, 10^2).$$

The independent error terms ε_{c^*sxt} in the test country's adjustment factor are Normally distributed with fixed variance,

$$\varepsilon_{c^*sxt} \sim \text{Normal}(0, 0.75^2).$$

The variance term σ_μ in Equation 3 is fixed at $\sigma_\mu = 0.5$ for this analysis, which controls the degree to which the adjusted mortality curves conform to the shape of those of other countries.

3.2.8 Computation

Parameters and estimates were produced using Hamiltonian Monte Carlo (HMC) implemented in `cmdstanr` (Gabry & Češnovar, 2021). Standard convergence checks including visual traceplot inspection and checking $\hat{R}$ values were performed.

4 Results

In this section we first present the extracted mortality curves for each of the countries of interest, and show the estimated adjustment factors for the training countries and estimates of the underlying parameters. In Sections 4.3 and 4.4, we present applications of the mortality adjustment process for Finland and the United States respectively. In Finland, where high quality data are available for some of the study period, we compare our adjusted mortality rates to assess performance of the method. For the United States we produce estimates of mortality rates where, to our knowledge, no reliable estimates exist for the period.

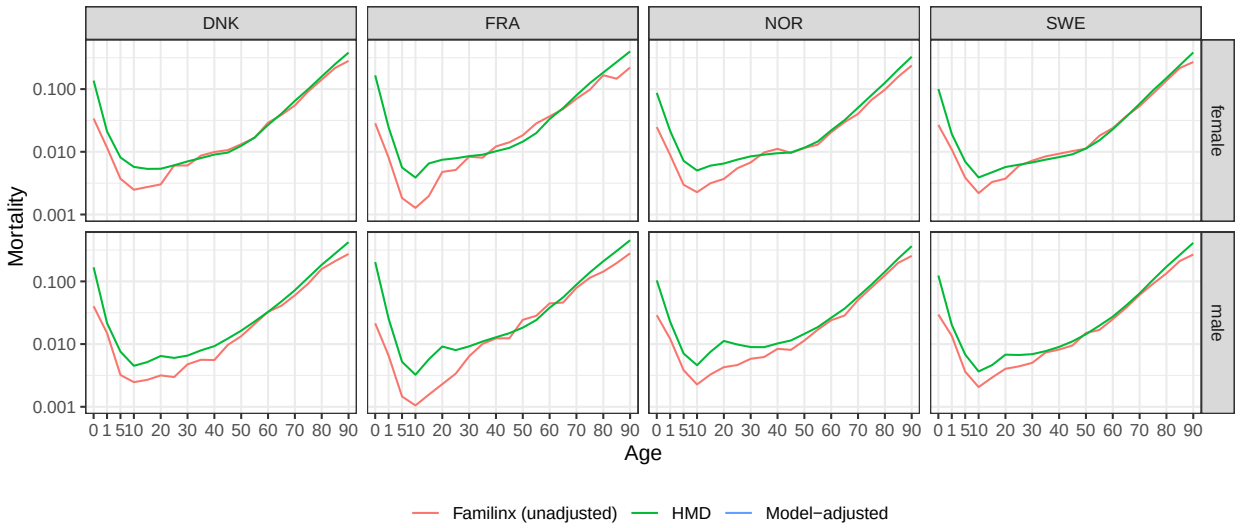


Figure 2: Comparison of mortality curves inferred from the Familinx data with those of the HMD for 1890 Denmark, France, Norway, and Sweden.

4.1 Extracted mortality curves

The implied sex-specific mortality curves from the genealogical data for the four training countries are shown in Figure 2 for the years 1895-1899, with HMD mortality curves shown where available. The set of curves for all years in the study period are given in the Appendix A.2.

Broadly speaking, the genealogy-derived mortality rates are lower than the HMD. This is particularly noticeable in infant and child mortality. For example, the HMD rate for the ages 5-10 1890 male population in France (0.0052 deaths per person-years lived) is more than triple the inferred genealogy rate (0.0015). The mortality rates typically become much closer in the mid-adult ages of around 40 to 65, and sometimes exceed the HMD rates, as seen in Swedish female population. At the oldest age groups, the genealogy-derived rates can sometimes dip downward, which is likely due to small counts leading to unstable mortality calculations.

4.2 Estimated adjustment factors

In Figure 3 we show the estimates of the adjustment factor ψ from the model described in Section 3.2 for the years 1895-1899. Posterior medians with 80% credible intervals are shown with solid lines and shaded regions respectively. As mentioned previously in Section 3.2.1,

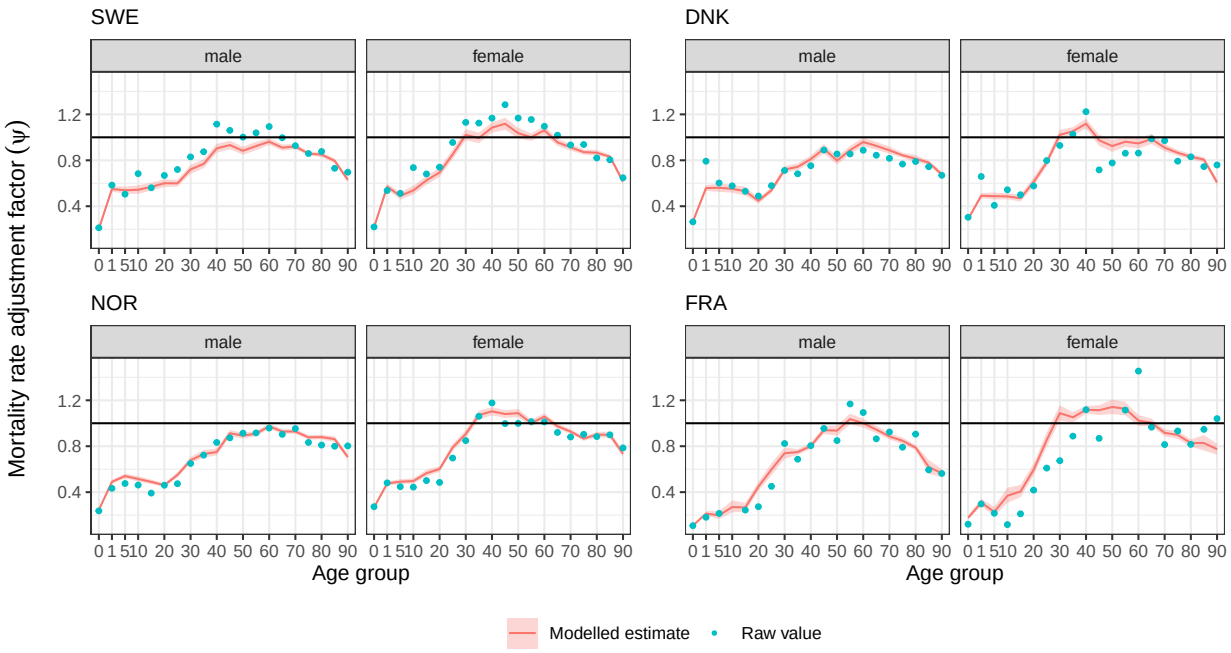


Figure 3: Estimated adjustment factors (ψ), representing the ratio of the genealogy-derived mortality rates to the HMD rates for 1890. Posterior medians with 80% credible intervals are shown by the red line and shaded area accordingly, and blue points represent raw values of mortality rate ratios.

the adjustment factor ψ represents the ratio of the Familinx mortality rate to the HMD rate. Therefore, values less than 1 indicate that the Familinx mortality rates are too low, whereas values greater than 1 indicate that Familinx rates are too high. Values of ψ for all training countries over the entire study period are shown in Appendix A.3

We can again observe a mortality underreporting bias in the younger ages, while rates are more representative in the adult ages. The structure and extent of bias varies between populations. Genealogy-derived rates for adult female populations appear closer to HMD rates than the corresponding male populations, despite the Familinx dataset being more heavily skewed towards males. There is also considerable variation between countries. For example, in the French male population, mortality reporting increases over age until peaking around age 55 before declining. Meanwhile in the Danish male population, mortality reporting is more uniform over age.

Mortality rate bias also evolves over time. Figure 4 shows the modelled estimates of ψ for female populations over 4 time periods, overlaid with the observed "raw" Familinx-HMD mortality rate ratio. Even within a country, the ratios can vary considerably between periods,

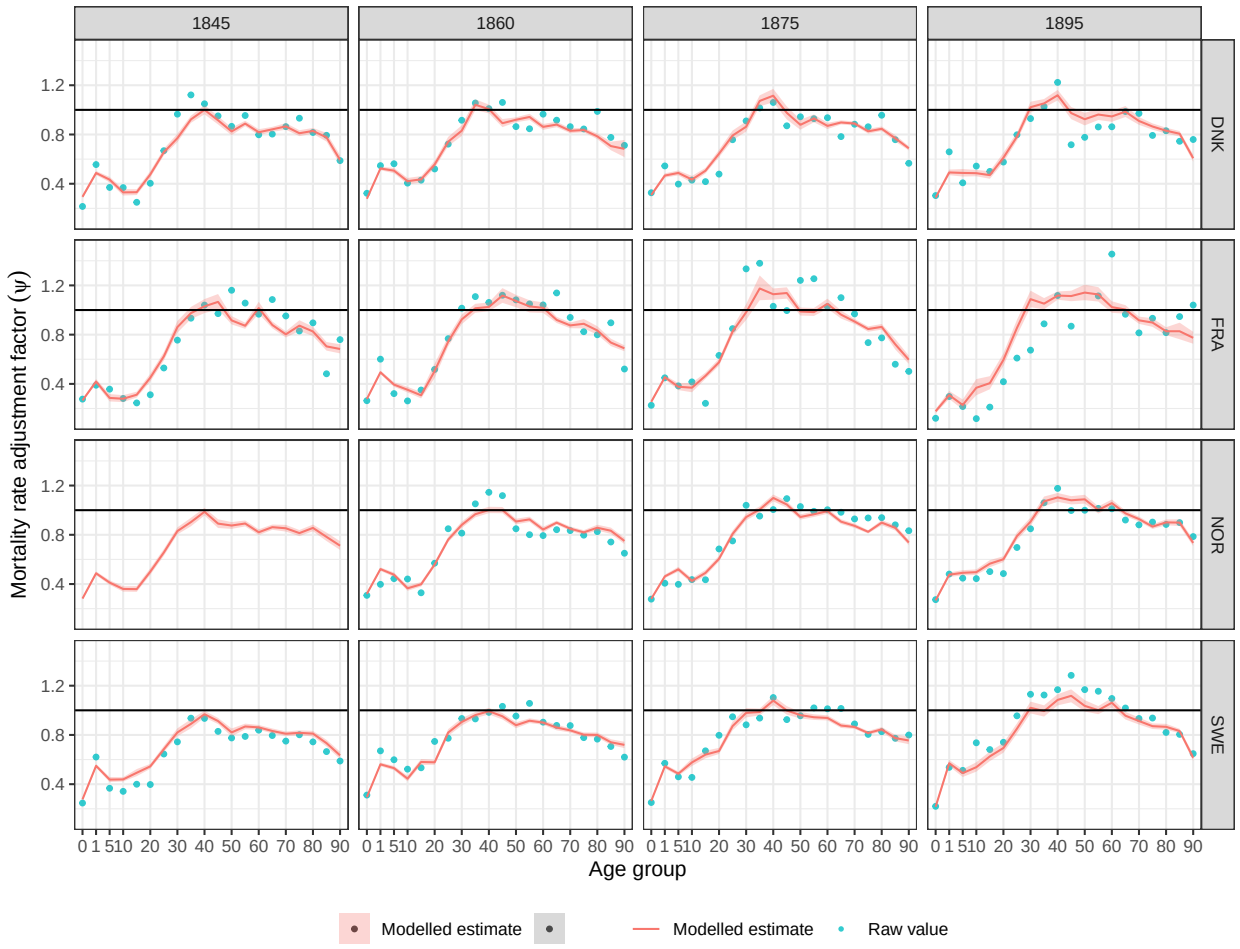


Figure 4: Estimated adjustment factors (ψ), representing the ratio of the genealogy-derived mortality rates to the HMD rates for the female populations over four select periods. Posterior medians with 80% credible intervals are shown by the red line and shaded area accordingly, and blue points represent raw values where available.

although it is difficult to distinguish any clear temporal trends. As expected, the modelled estimates are smoother than the raw ratios due to limited flexibility in the model.

The age effect α is shown in Figure 5, which more directly shows the mortality underreporting in younger ages. Lower values of α indicate more underreporting. Estimates of α increase with age, and are generally more stable after around age 45.

One source of temporal variation in the modelled estimates is the cohort effect γ, which is shown in Figure 6. Compared to the age effect α, the overall trends over cohort are much smaller. The point estimates in the right panel of Figure 6 suggest that (after accounting for the covariates) mortality reporting slightly improves over time until around the 1870 cohort,

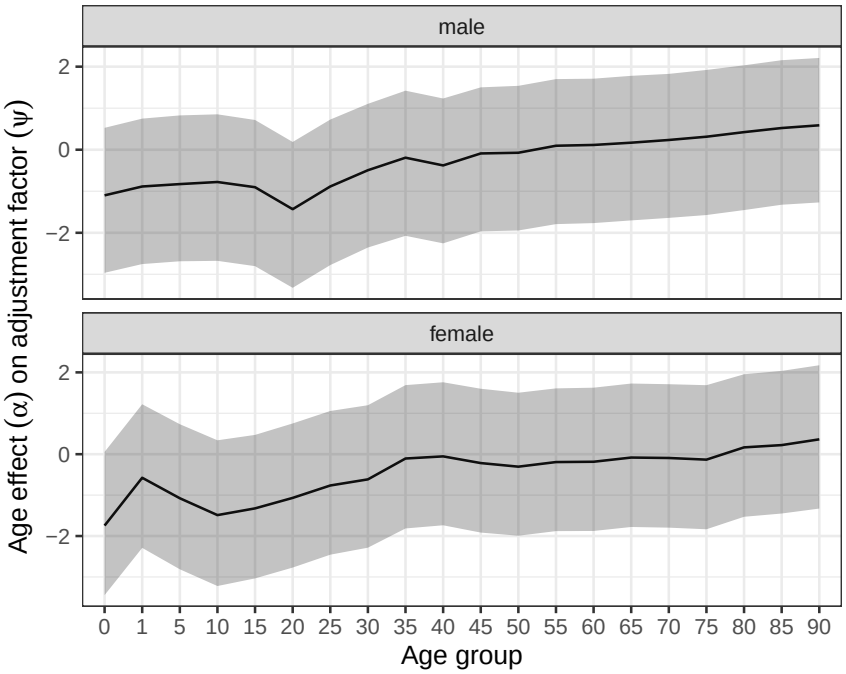


Figure 5: Estimated age effects α on the adjustment factor ψ. The solid black line represents the posterior median estimate, and the grey shaded region represents an 80% credible interval.

where there is a small decline. This may be explained with the choice of Kaplanis et al. (2018) to remove the profiles of living individuals from the Familinx data to protect their privacy.

4.3 Application to Finland

Adjusted mortality rates can be produced by dividing the raw genealogy-derived rate by the adjustment factor ψ. To evaluate the performance of this procedure, we compare our adjusted mortality curves to more reliable estimates from the HMD. HMD estimates are available in Finland from 1878 onward, which allows for comparison over the last five periods spanning 1875 to 1899. Figure 7 compares the raw genealogy rate, HMD rate, and model-adjusted rates for 1895-1899 Finland. Mortality curves for the entire study period are given in Appendix A.4. The model typically adjusted the Familinx rates upward for the young ages. There is less of an adjustment in the older age groups where the Familinx and HMD rates usually agree more closely. Occasionally the model can overcorrect, resulting in mortality rate estimates higher than the HMD.

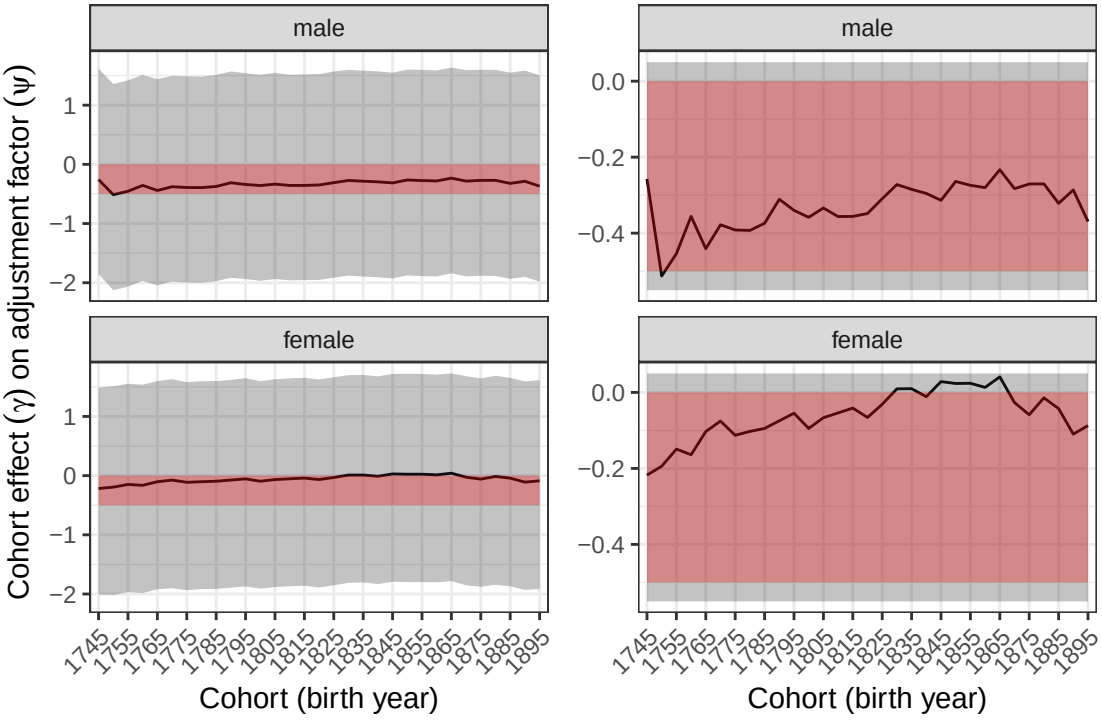


Figure 6: Estimated cohort effects on the adjustment factor ψ. The right panel shows the estimate on a magnified scale. The posterior median of the cohort effect γ is shown with a solid black line, and the grey shaded region represents an 80% credible interval. The red shaded region acts only as a reference between the two panels.

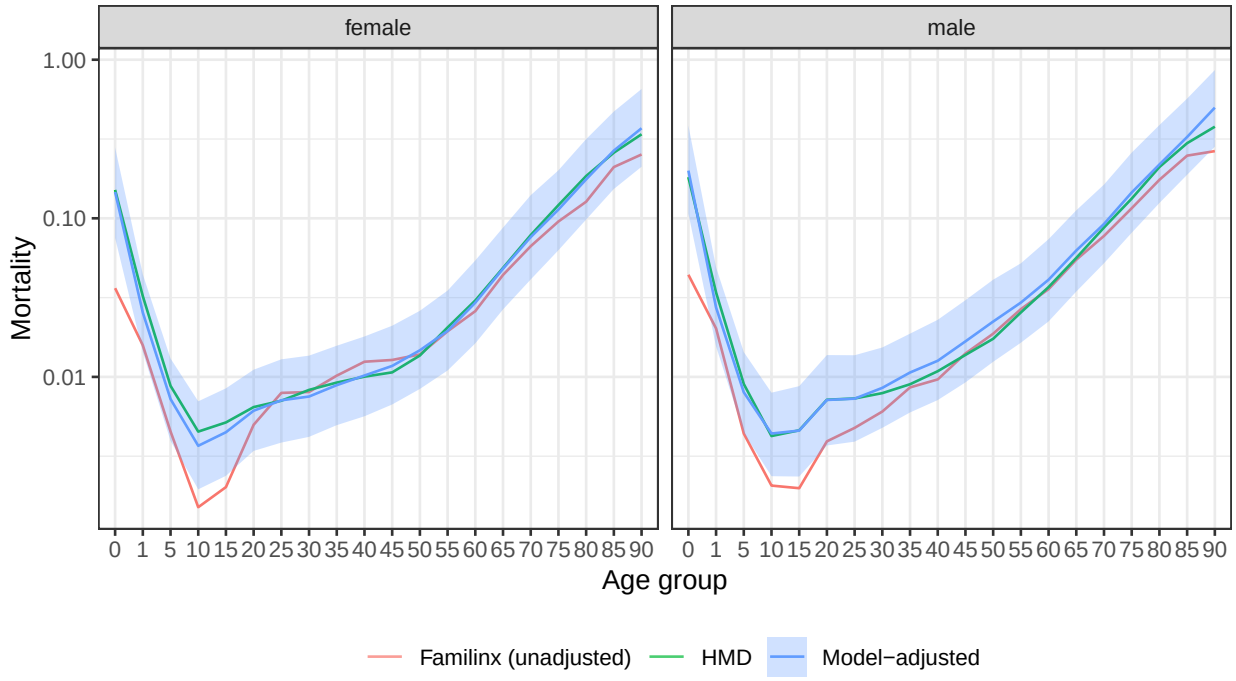


Figure 7: Mortality estimates for 1895 Finland before and after model adjustment. The HMD rate (green) for this period is shown for reference. For the model-adjusted mortality estimate, the shaded blue region represents an 80% credible interval.

Figure 8 shows estimates for the life expectancies at birth in Finland over the entire estimation period. The resulting life expectancies at birth are lower and much closer to values from the HMD. For the male population, overcorrected mortality rates result in life expectancies that are too low. The dip in life expectancy in the 1865-1869 period may be explained by the Great Finnish Famine of 1866-1888.

4.4 Application to the United States

Unlike Finland, no reliable mortality estimates exist for this period in the United States. Here we apply our procedure to produce estimates age-specific mortality and life expectancy from 1835-1900. Results from applying the bias correction are shown in Figure 9. Mortality rates are adjusted for younger age groups, but are relatively unchanged for older adult groups. The full set of mortality curves over the entire study period are given in Appendix A.5. Life expectancies at birth over the estimation period are shown in Figure 10. The drop in life expectancy in the 1860-1864 period in the male population coincides with the US Civil War. To our knowledge, no nationally representative high quality mortality data exist for this

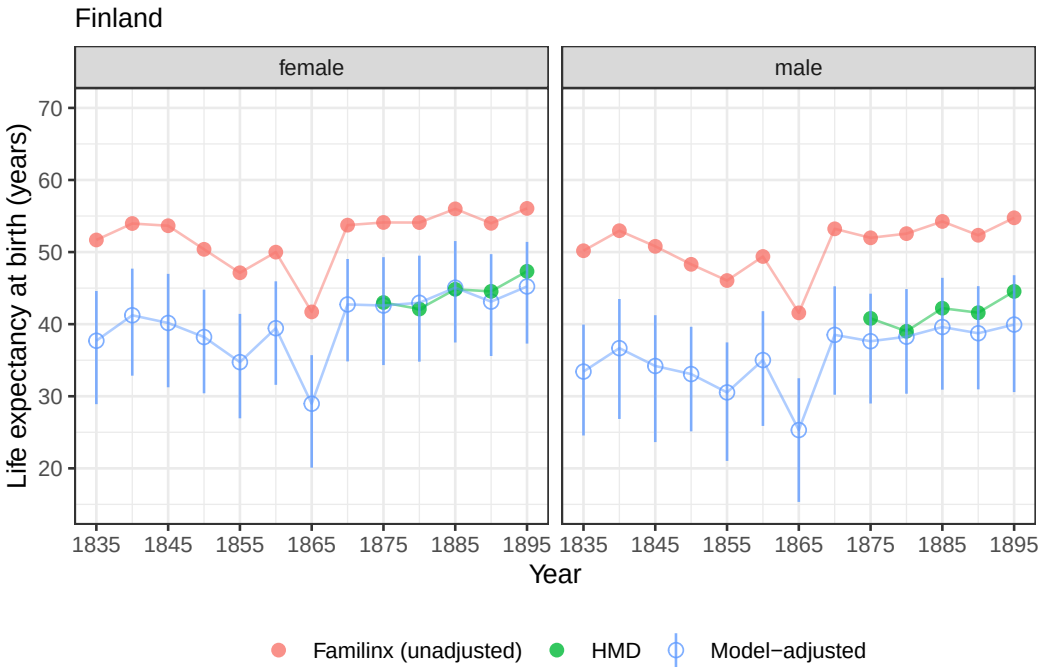


Figure 8: Life expectancies at birth for 1835-1899 Finland before and after model adjustment. Estimates from HMD rates (green) are shown for reference. Error bars represent an 80% credible interval.

period, although one reference data point is provided by Glover (1921) based on historical life tables in the state of Massachusetts. Life expectancy at birth given by Glover (1921) is lower than the model-adjusted estimate which suggests the need for further adjustment.

5 Discussion

In this study we demonstrate that granular mortality rates inferred from the Familinx genealogical dataset can be misrepresentative when compared against a more reliable data source. Age specific mortality rates inferred from the genealogy are typically too low for younger ages, but are more representative in mid- and older adult ages.

Using a statistical model to estimate the mortality rate bias in countries with data overlap, we also propose a method to produce adjusted genealogy-derived mortality rate estimates for country-periods where reliable data is not available. The statistical model allows the bias to vary by age and cohort, and over a set of data quality indicators. An additional prior is placed on the age structure of the adjusted mortality estimates to smooth the mortality

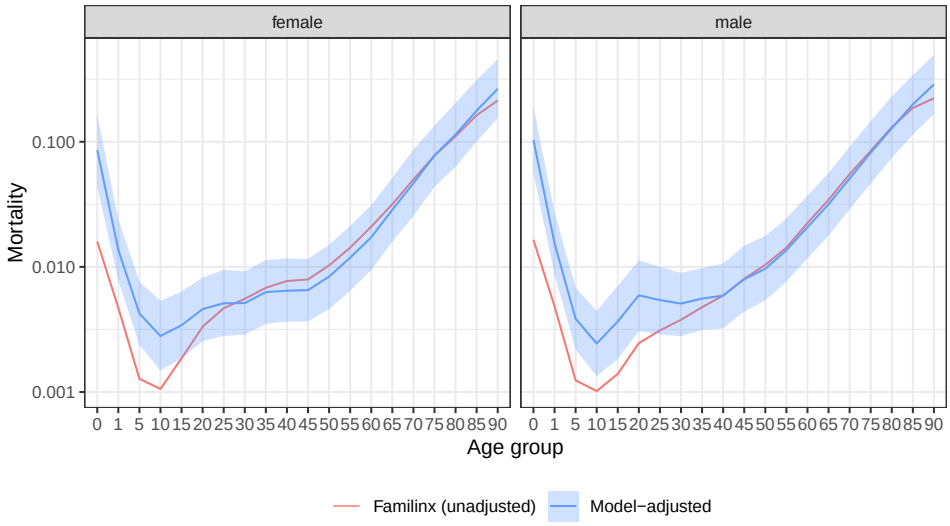


Figure 9: Mortality estimates for 1890 United States before and after model adjustment. The HMD rate for this period is shown for reference. For the model-adjusted estimate, the blue shaded region represents an 80% credible interval.

Figure 10: Life expectancies at birth for 1835-1899 United States before and after model adjustment. Error bars represent an 80% credible interval.

curve over age and constrain estimates to be roughly similar in shape to other observed mortality curves. We apply our method to estimate age-specific mortality in Finland and the United States over the period 1835-1899. In both cases, model-based mortality rate adjustments result in lower life expectancies at birth compared to the raw genealogy-based estimate. The adjusted life expectancies for Finland agree closely with held-out data from Human Mortality Database where data exist over the period 1878-1899.

The apparent bias in the Familinx-inferred mortality 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, Kaplanis et al. (2018) only include profiles which contained exact birth and death dates to avoid age heaping, which also may have resulted in an overall higher quality sample. 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 (calculated as the number of deaths over the exposure-to-risk, or person-years lived) 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 the analysis.

The resulting estimates are also sensitive to model choices. One particularly influential choice is the amount of variation allowed from the "expected shape" in Equation 3 controlled by σ_μ^2. The modeller exercises discretion on how closely the test country mortality should conform to that of the training countries. For example, we might expect that Finland is quite similar to other Nordic countries and therefore choose a smaller value of σ_μ^2. Meanwhile for the United States, a larger value of σ_μ^2 may be desirable to allow for more variation. In particular, during periods where there are acute changes to mortality such as during the Civil War, overly constraining the mortality curve might diminish legitimate patterns in the data.

One important assumption in our calculation of mortality rates from these data is that of no international migration by only assigning each individual to a single country. Since the individual profiles contain only birth and death information, the exposure-to-risk from international migrants are misattributed for some portion of their life. This likely dispropor-

tionately affects the exposure-to-risk in high-migration countries like the United States. For example, if a country experiences high immigration, then under this scheme the exposure will be overestimated particularly for younger ages (and the mortality consequently underestimated) since their exposure prior to migration is being attributed to the destination country. To adjust exposure for migration, one could possibly incorporate more complex assumptions and estimate age at migration based on birthplaces of children, and/or count individuals with different birth and death countries to estimate the proportion of migrants in the population.

In light of these limitations, there are many avenues for future research. One potential strategy is to incorporate additional available data sources. This could involve incorporating additional online or offline genealogical databases (Smith & Mineau, n.d.), or using subnational data as gold standard data if national data are not available.

Furthermore, it is not yet clear how demographic processes and online user patterns contribute to the observed bias. For instance, the mortality underreporting observed in younger ages may be a result of lack of descendants, but another explanation could be the preferential use of genealogical tools by (descendants of) low mortality populations. The picture is further confounded by potentially different birth rates and rates of migration in high and low mortality populations, not to mention the fact that not all individuals leave a paper trace (Hollingsworth, 1976). Disentangling these contributing factors, possibly by analyzing the present-day user base may help us characterize which populations are underrepresented, and therefore help in more directly adjusting for the bias.

We focused on correcting the mortality rates implied by online genealogies to reconstruct life expectancy, a widely used measure. Our approach can be extended to study other big questions in historical demography. Online genealogies may be used to characterize kinship transitions on a global scale - how do family structures and kin availability change throughout the demographic transition? Crowd-sourced data could also be used to reconstruct historical migration patterns to outline, for example, processes of urbanization. These are just two examples of how genealogies can help overcome historical data limitations and provide new insights into large-scale social phenomena in a comparative perspective. Our work constitutes a first step towards making online genealogies more useful by improving the reliability of the measures we can derive from them.

References

Alburez-Gutierrez, D., Kolk, M., & Zagheni, E. (2021, October). Women's Experience of Child Death Over the Life Course: A Global Demographic Perspective. *Demography*, *58*(5), 1715–1735. Retrieved 2021-10-20, from https://read.dukeupress.edu/demography/article/58/5/1715/174263/Women-s-Experience-of-Child-Death-Over-the-Life doi: 10.1215/00703370-9420770

Alexander, M., Polimis, K., & Zagheni, E. (2020, August). Combining Social Media and Survey Data to Nowcast Migrant Stocks in the United States. *Population Research and Policy Review*. Retrieved 2020-09-17, from http://link.springer.com/10.1007/s11113-020-09599-3 doi: 10.1007/s11113-020-09599-3

Blanc, G. (2020). Modernization Before Industrialization: Cultural Roots of the Demographic Transition in France. *SSRN Electronic Journal*. Retrieved 2021-05-31, from https://www.ssrn.com/abstract=3702670 doi: 10.2139/ssrn.3702670

Fire, M., & Elovici, Y. (2013, November). Data Mining of Online Genealogy Datasets for Revealing Lifespan Patterns in Human Population. *arXiv:1311.4276 [cs, q-bio, stat]*. Retrieved 2019-01-02, from http://arxiv.org/abs/1311.4276 (arXiv: 1311.4276)

Gabry, J., & Češnovar, R. (2021). cmdstanr: R interface to 'cmdstan' [Computer software manual]. (https://mc-stan.org/cmdstanr, https://discourse.mc-stan.org)

Glover, J. W. (1921). *United states life tables, 1890, 1901, 1910, and 1901-1910: Explanatory text, mathematical theory, computations, graphs, and original statistics, also tables of united states life annuities, life tables of foreign countries, mortality tables of life insurance companies*. US Government Printing Office.

Henry, L. (1956, April). Anciennes familles genevoises. Etude démographique: XVIme - XXme siècle. *Population (French Edition)*, *11*(2), 334. Retrieved 2021-05-31, from https://www.jstor.org/stable/1524668?origin=crossref doi: 10.2307/1524668

Holden, L., & Boudko, S. (2018, September). The Norwegian Historic Population Register and Migration. *Journal of Migration History*, *4*(2), 249–263. Retrieved 2020-09-17, from https://brill.com/view/journals/jmh/4/2/article-p249_249.xml doi: 10.1163/23519924-00402002

Hollingsworth, T. (1976). Genealogy and historical demography. *Annales de démographie historique*, *1976*(1), 167–170. Retrieved 2021-06-16, from https://www.persee.fr/doc/adh_0066-2062_1976_num_1976_1_1310 doi: 10.3406/adh.1976.1310

Kaplanis, J., Gordon, A., Shor, T., Weissbrod, O., Geiger, D., Wahl, M., ... Erlich, Y. (2018, April). Quantitative analysis of population-scale family trees with millions of relatives.

Science, *360*(6385), 171–175. Retrieved 2019-01-03, from http://www.sciencemag.org/lookup/doi/10.1126/science.aam9309 doi: 10.1126/science.aam9309

Kolk, M. (2014, January). Multigenerational transmission of family size in contemporary Sweden. *Population Studies*, *68*(1), 111–129. Retrieved 2018-11-27, from http://www.tandfonline.com/doi/abs/10.1080/00324728.2013.819112 doi: 10.1080/00324728.2013.819112

Murphy, M. (2004, May). Tracing very long-term kinship networks using SOCSIM. *Demographic Research*, *10*, 171–196. Retrieved 2021-05-31, from http://www.demographic-research.org/volumes/vol10/7/ doi: 10.4054/DemRes.2004.10.7

Murphy, M. (2011, January). Long-Term Effects of the Demographic Transition on Family and Kinship Networks in Britain. *Population and Development Review*, *37*, 55–80. Retrieved 2019-08-06, from http://doi.wiley.com/10.1111/j.1728-4457.2011.00378.x doi: 10.1111/j.1728-4457.2011.00378.x

Oeppen, J. (2002, May). Enhanced: Broken Limits to Life Expectancy. *Science*, *296*(5570), 1029–1031. Retrieved 2019-08-30, from http://www.sciencemag.org/cgi/doi/10.1126/science.1069675 doi: 10.1126/science.1069675

Smith, K. R., & Mineau, G. P. (n.d.). The Utah population database: The legacy of four decades of demographic research. *Historical Life Course Studies*.

van Raalte, A. A., Sasson, I., & Martikainen, P. (2018, November). The case for monitoring life-span inequality. *Science*, *362*(6418), 1002–1004. Retrieved 2021-01-15, from https://www.sciencemag.org/lookup/doi/10.1126/science.aau5811 doi: 10.1126/science.aau5811

Verdery, A. M. (2015, September). Links Between Demographic and Kinship Transitions. *Population and Development Review*, *41*(3), 465–484. Retrieved 2021-03-25, from http://doi.wiley.com/10.1111/j.1728-4457.2015.00068.x doi: 10.1111/j.1728-4457.2015.00068.x

Wrigley, E. A., & Schofield, R. S. (1983, July). English Population History from Family Reconstitution: Summary Results 1600-1799. *Population Studies*, *37*(2), 157. Retrieved 2021-05-31, from https://www.jstor.org/stable/2173980?origin=crossref doi: 10.2307/2173980

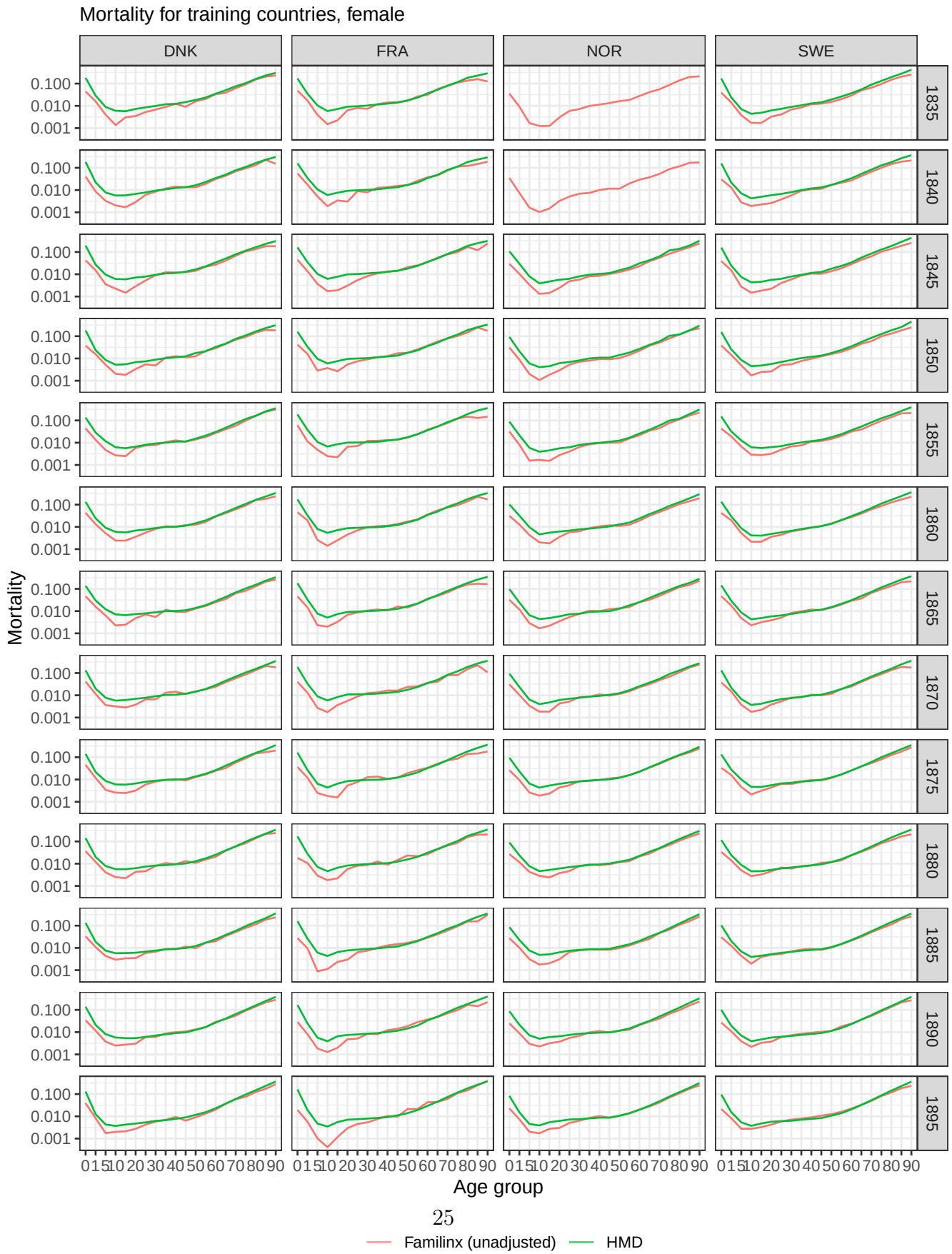
Zhao, Z. (2001, January). Chinese genealogies as a source for demographic research: A further assessment of their reliability and biases. *Population Studies*, *55*(2), 181–193. Retrieved 2020-09-17, from http://www.tandfonline.com/doi/abs/10.1080/00324720127690 doi: 10.1080/00324720127690

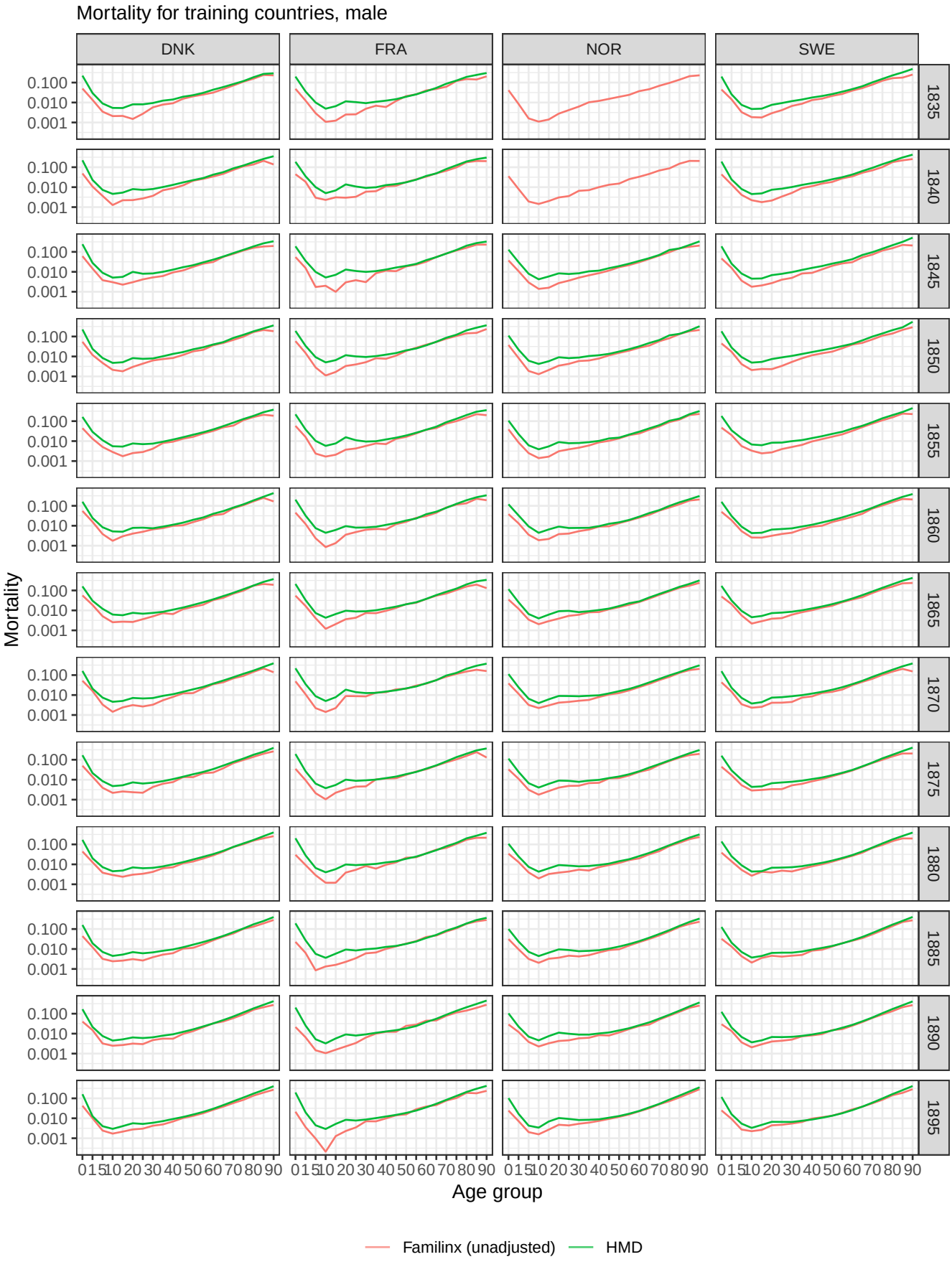
A Appendix

A.1 Strings used for country assignment

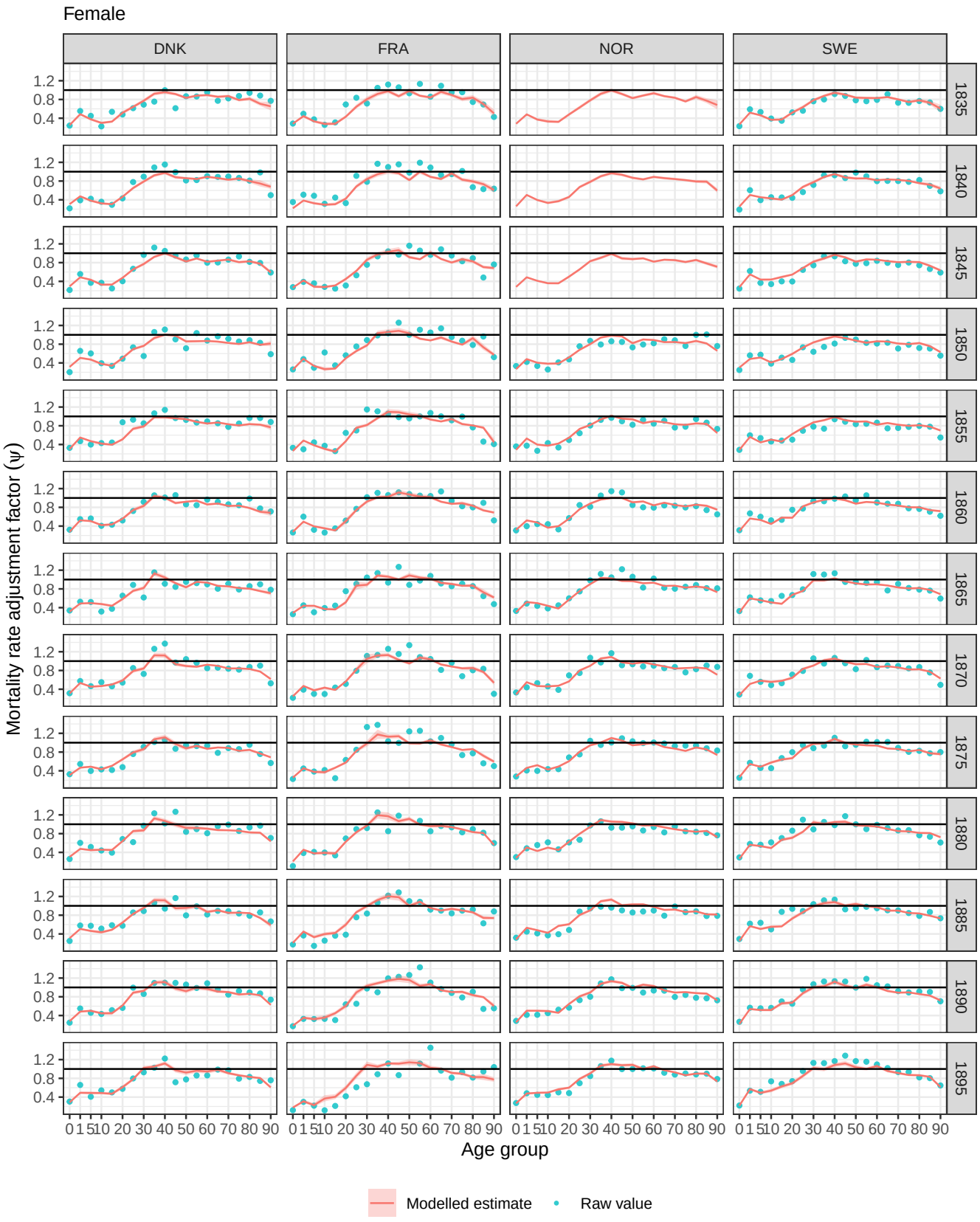
Country	Strings used
USA	', US$', 'USA', 'UNITED STATES', 'UNITED STATES OF AMERICA', ', AMERICA', 'REPUBLIC OF AMERICA', 'NEW ENGLAND', 'UNITED COLONIES OF AMERICA', 'MASSACHUSETTS', 'CONNECTICUT', 'ILLINOIS', 'NEW YORK', 'NORTH CAROLINA', 'SOUTH CAROLINA', 'NEW HAMPSHIRE', 'KENTUCKY', 'RHODE ISLAND', 'COLONIAL AMERICA', 'PROVINCE OF VIRGINIA'
FRA	', FRA$', 'FR$', 'FRANCE', 'FRANCIA', 'FRANKRIKE', 'FRANKREICH', 'FRANKRÄICH'
FIN	' FIN$', 'FI$', 'FINLAND', 'SUOMI', 'SOOMLANE', 'FINLANDIA', 'FINLANDE'
DNK	'DK$', 'DENMARK', 'DANMARK', 'DANEMARK', 'TANSKA', 'DÄNEMARK', 'DINAMARCA', 'DNK'
NOR	' NOR$', ' NO$', 'NORWAY', 'NORWEGEN', 'NORGE', 'NORUEGA', 'NORJA', 'NORVÈGE'
SWE	' SWE$', ' SE$', 'SWEDEN', 'SVERIGE', 'SCHWEDEN', 'SUECIA', 'RUOTSI', 'SUÈDE', 'SUÃ¨DE'

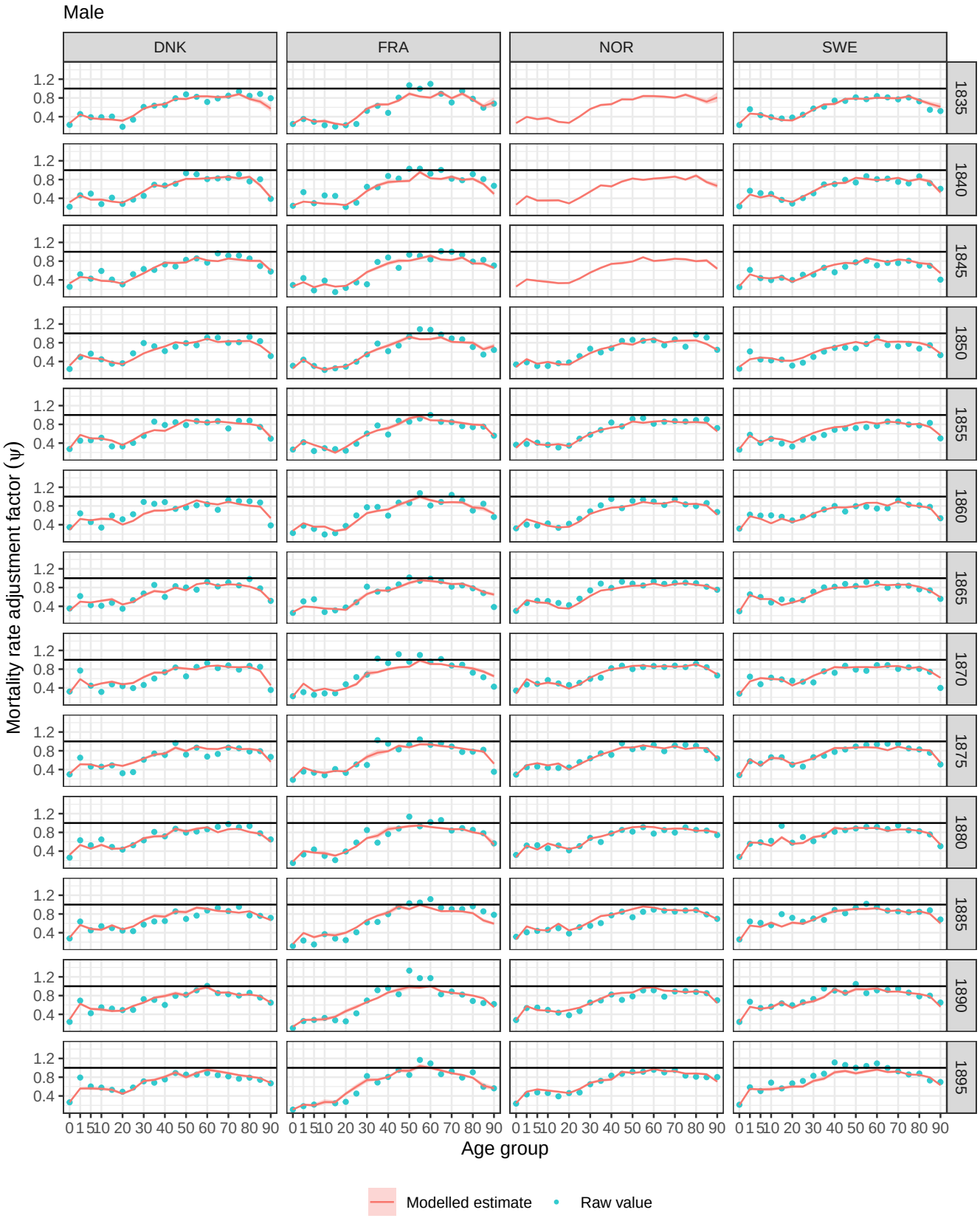
A.2 Extracted mortality for all training countries



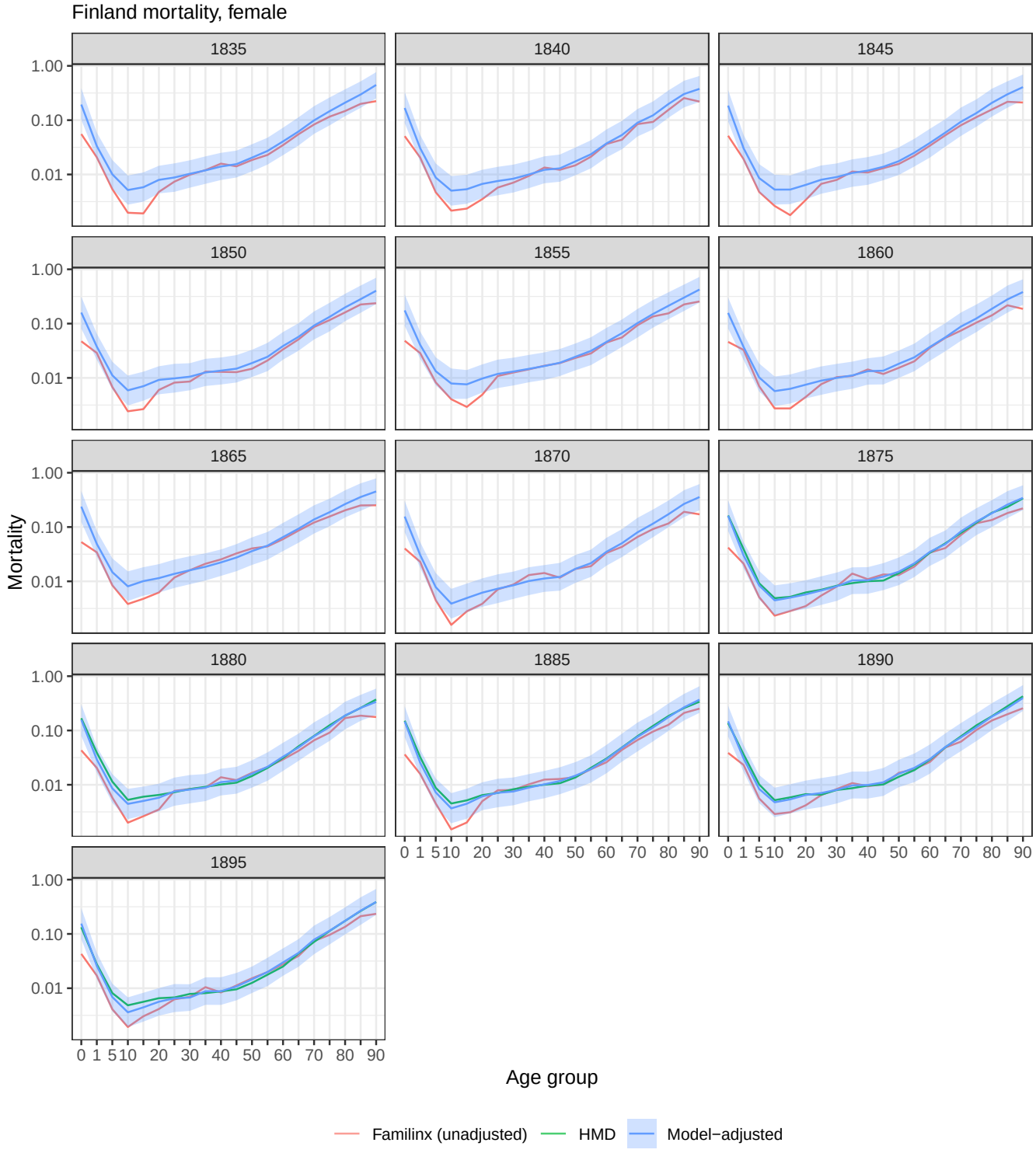

Mortality for training countries, male
DNK
FRA
NOR
SWE
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
Mortality
0.100
0.010
0.001
0 1 5 10 20 30 40 50 60 70 80 90
Age group
Familinx (unadjusted)
HMD

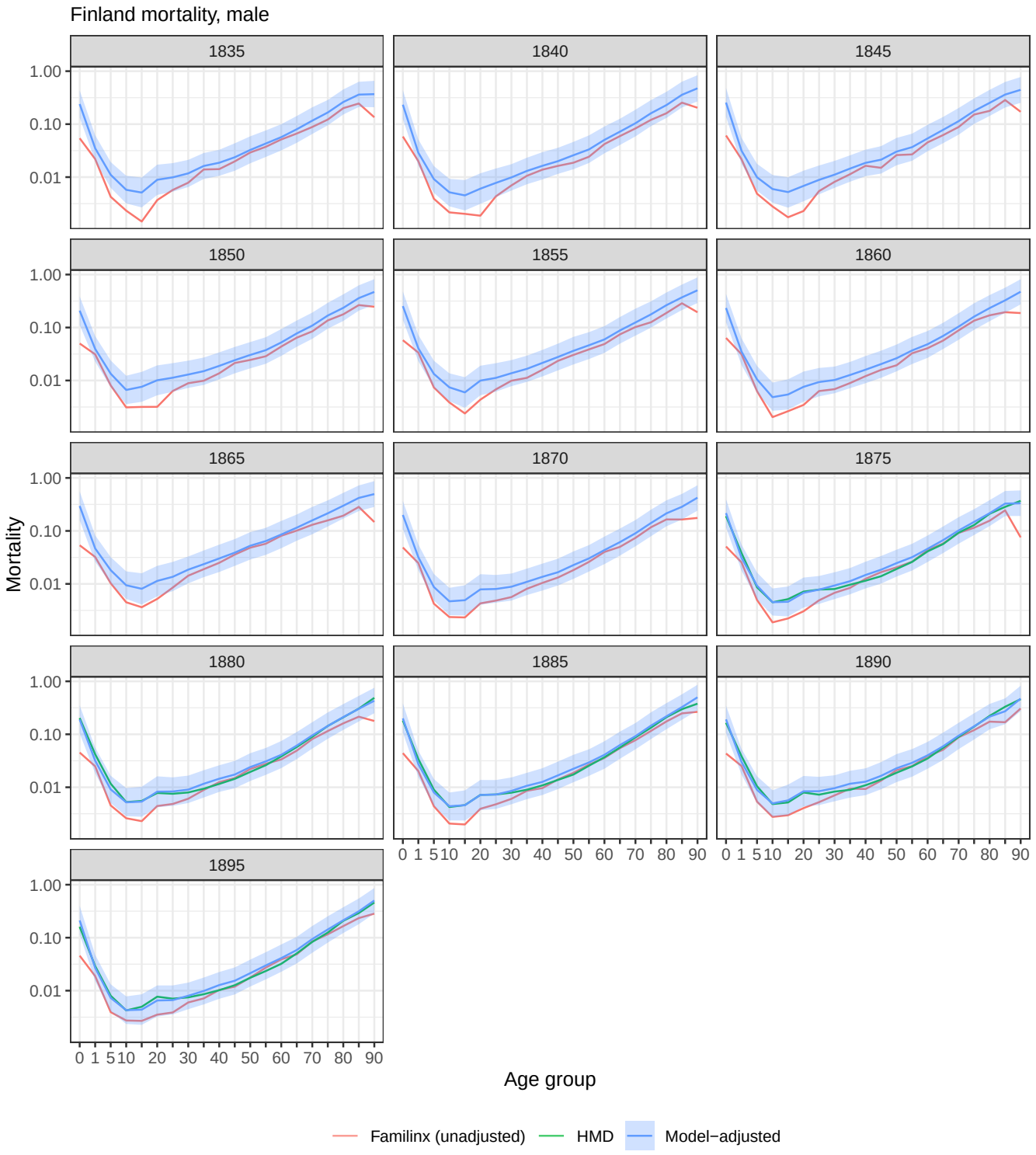
A.3 Estimated adjustment factors for all training countries



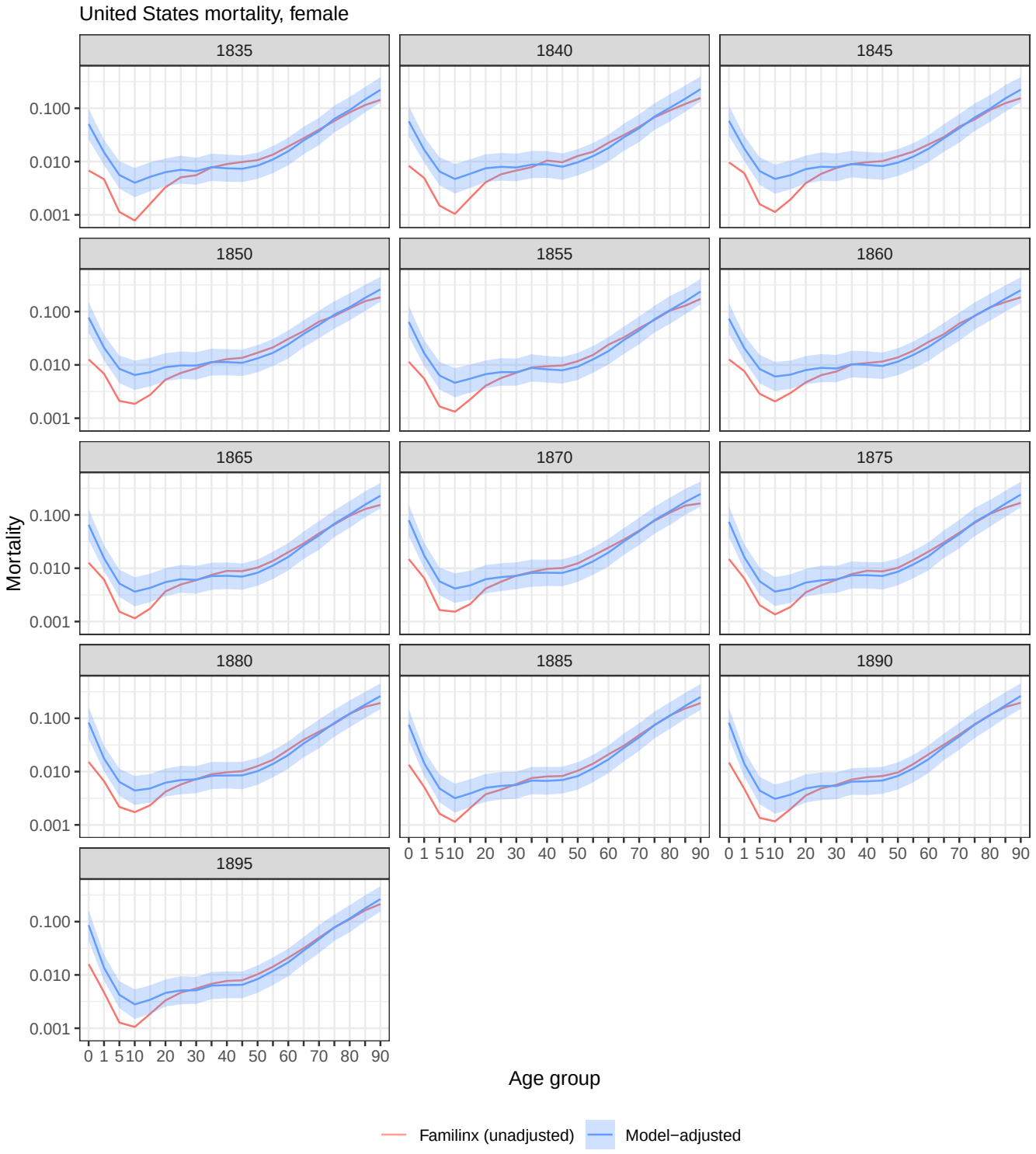

Male
DNK
FRA
NOR
SWE
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1.2
0.8
0.4
Mortality rate adjustment factor (ψ)
0 1 5 10 20 30 40 50 60 70 80 90
Age group
Modelled estimate
Raw value

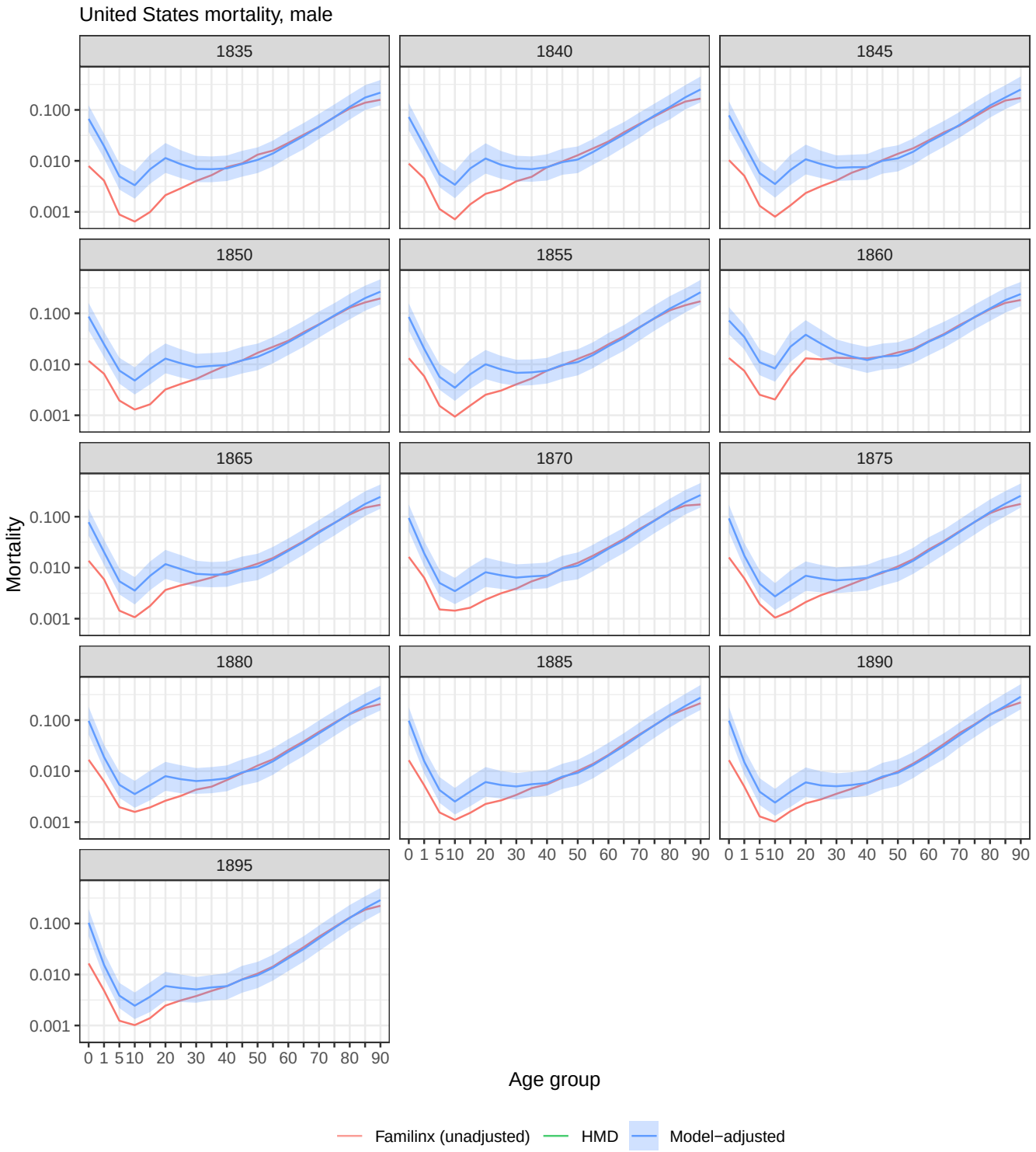
A.4 Adjusted mortality curves for Finland 1835-1899




Finland mortality, male
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
1.00
0.10
0.01
Mortality
0 1 5 10 20 30 40 50 60 70 80 90
Age group
Familinx (unadjusted)
HMD
Model−adjusted

A.5 Adjusted mortality curves for the United States 1835-1899




United States mortality, male
1835
1840
1845
1850
1855
1860
1865
1870
1875
1880
1885
1890
1895
0.100
0.010
0.001
Mortality
0 1 5 10 20 30 40 50 60 70 80 90
Age group
Familinx (unadjusted)
HMD
Model−adjusted

A.6 Right singular vectors of log mortality rates