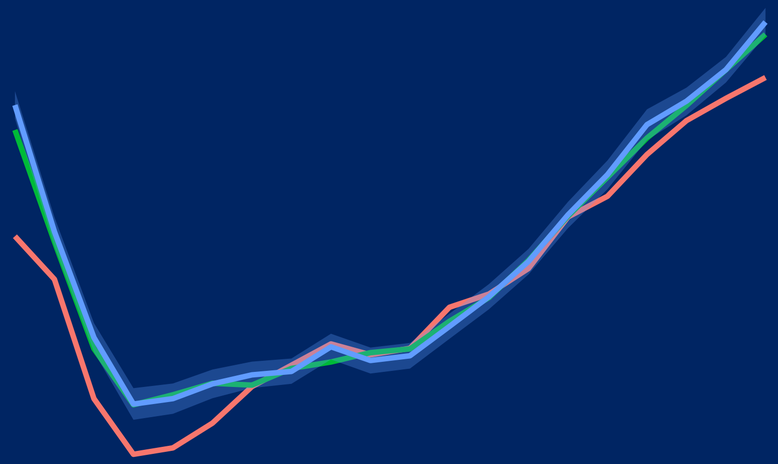


Identifying and Correcting Bias in Big Crowd-Sourced Online Genealogies

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

[1] University of Toronto
[2] Max Planck Institute for Demographic Research
* myc.chong@mail.utoronto.ca

PAA 2021 Annual Meeting - Flash: New Research in Historical Demography - May 7, 2021

Online genealogies and demographic research

Background

Online genealogical data is a rich data source, containing detailed information about life courses and family structure. The Familinx genealogical dataset [1] is one prominent example, and indeed holds potential for making new inferences about historical populations.

However, these data are subject to inaccuracies, underreporting, and biased representation of certain groups.

Objective

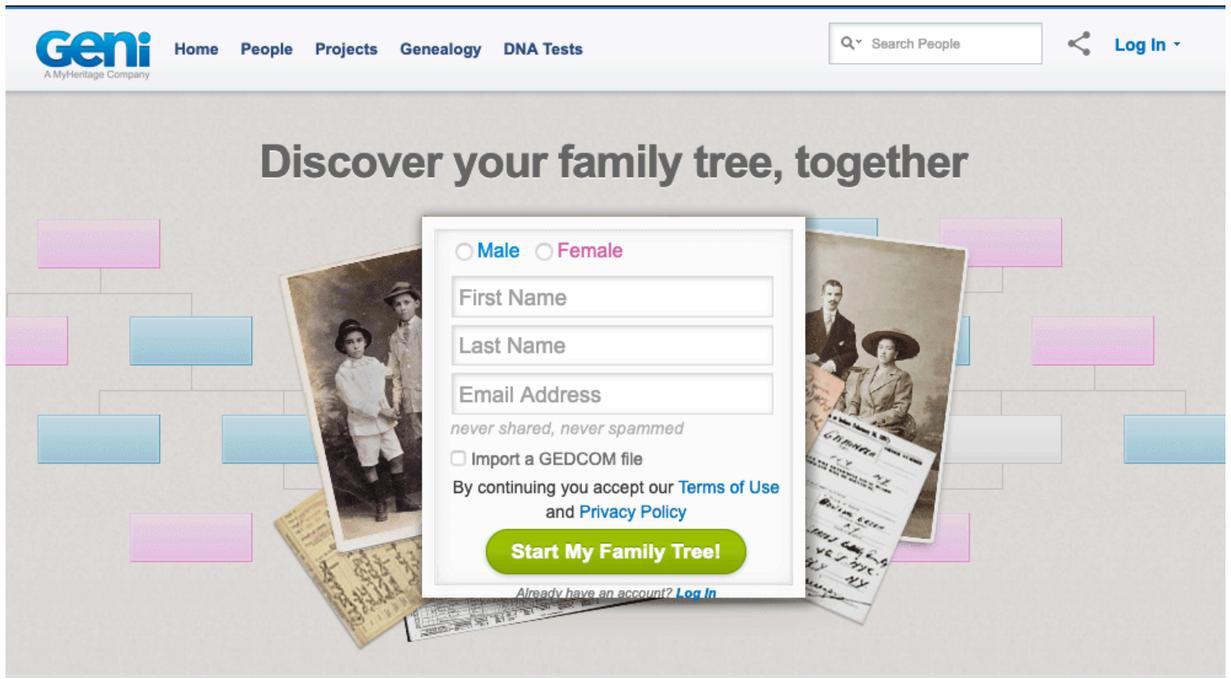
In this study, we investigate biases in the mortality rates derived from the Familinx dataset, and present a statistical model to correct mortality rates by calibrating against a more reliable data source.

The Familinx dataset

The Familinx dataset contains information on over 86 million individuals, aggregated from family trees created by users of Geni.com.

Data fields include date and place of birth, date and place of death, and kin relationships.

Data quality varies by time period and country. Profiles are largely concentrated in Europe and North America.



Data issues in online genealogies

The production of genealogies is contingent on historical and social forces. As a result, the resulting dataset may suffer from non-representativeness issues such as [2]:

- **gender bias** - women are underrepresented
- **ascendant genealogies** - those without descendants may be less likely to be recorded
- **survivor bias** - early deaths may not be recorded
- **systematic discrepancies in access to recorded family histories**

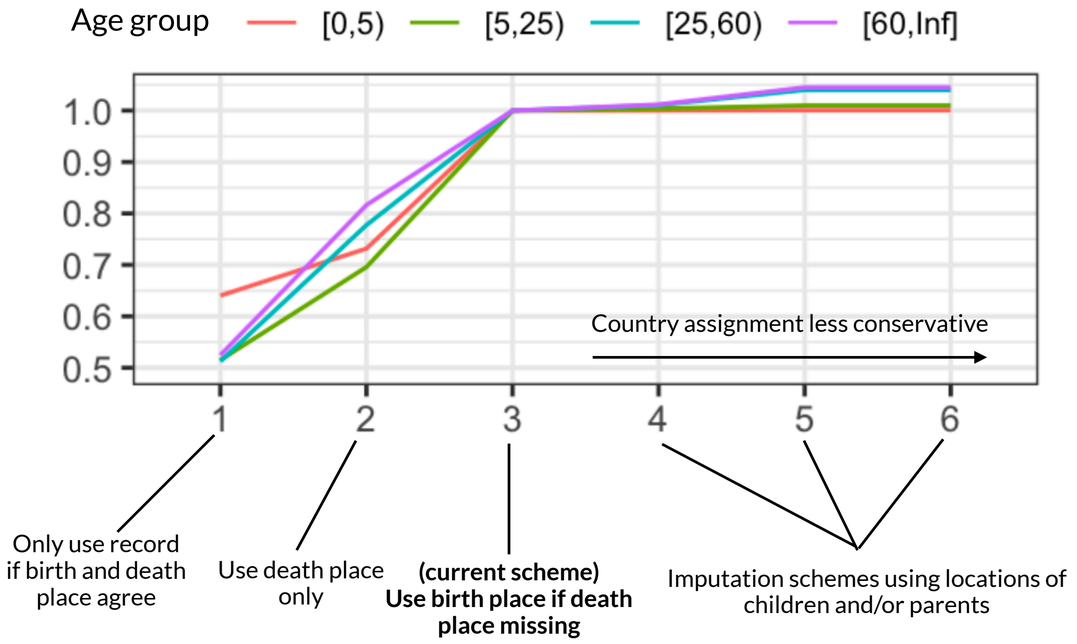
It is therefore critical that researchers using online genealogies for demographic inferences understand what biases may be present in the data, how they manifest in population-level estimates, and account for them when possible.

Extraction of demographic information

Partially missing data, incorrect records, and free-text location names mean that **extracting demographic information is not straightforward**. Past profiles also do not contain information about place of residence apart from possibly vital event locations. For our countries of interest, we infer the place of residence and place of death for an individual by matching location names in the relevant data fields. We make the following simplifying assumptions regarding an individual's location:

- Individuals do not migrate between countries
- An individual is assumed to reside and die in the country of their death, if recorded
- If the place of death is not recorded, then place of birth is used

We also investigated how many profiles were included under different assumptions about an individual's location. The plot below shows the relative number of deaths in the United States during the study period under different assumptions.



Mortality rates

Life tables and death rates are calculated using standard demographic procedures for **age groups 0, 1, 5, 10, …, 90+ in 5** year intervals from **1835 to 1900** by counting the deaths and aggregating the exposure for each age group for male and female populations in each period and country.

On the right are the mortality rates (top) for 1835 Sweden, and life expectancies (bottom) estimated from the Human Mortality Database (HMD)[3] and Familinx. Mortality rates inferred from the Familinx data are lower than the HMD, and the life expectancies at birth are consequently higher.

	Female	Male
HMD	44.3 years	39.9 years
Familinx	57.4 years	55.3 years

Modelling strategy

Our modelling strategy is to **calibrate the Familinx mortality rates using reliable, high quality data** from the HMD where available, then **produce adjusted mortality estimates** where HMD estimates are unavailable. We apply our method to the period 1835-1900 to take advantage of four countries where historical mortality rates are available. Finland is used as a test case since HMD estimates are only partially available in this time period, and we illustrate use of our method for the United States, where high quality data is not available.

“Training” cases - HMD available 1835-1900

- Denmark
- Sweden
- France
- Norway (available 1846 onward)

“Test” cases:

- Finland (HMD available from 1878 onward)
- United States - high quality data not available

Model description

The difference between the Familinx mortality rate μ and HMD mortality rate M is captured by an adjustment factor ψ.

For country c, sex s, age group x, and year t, and cohort $b(x, t)$, the mortality is modelled as:

Familinx deaths process	$d_{csxt} \sim \text{Poisson}(p_{csxt}\,\mu_{csxt})$	Familinx deaths d_{csxt} are modelled as a Poisson process with some known exposure p_{csxt} and rate μ_{csxt}.
Smoothing mortality rates	$\log \mu_{csxt} = V_{csx}\nu_{csxt} + \varepsilon_{csxt}$	To smooth the Familinx mortality rates, the log-rate $(\log \mu_{csxt})$ is expressed using the first 3 principal components of the singular value decomposition of the country's rates.
Relationship to HMD	$\log M_{csxt} \sim \text{Normal}\left(\log\left(\frac{\mu_{csxt}}{\psi_{csxt}}\right), \sigma_\psi^2\right)$	The (log) HMD mortality rate is related to the Familinx rate μ_{csxt} via an adjustment factor ψ_{csxt}, which can be interpreted as the ratio of the Familinx rate to the HMD rate.
Structure of relationship	$\log \psi_{csxt} = \alpha_{sx} + \gamma_{csb(x,t)} + X_{csxt}\beta_{sx}$ $\gamma_{csb(x,t)} \sim \text{Normal}(\bar{\gamma}_{sb(x,t)}, \sigma_\gamma^2)$	The adjustment factor ψ_{csxt} is allowed to vary with an age effect α, a hierarchical cohort effect γ, and a set of covariates X.

Data quality indicators

$$\log \psi_{csxt} = \alpha_{sx} + \gamma_{csb(x,t)} + X_{csxt}\beta_{sx}$$

We define a set of covariates X_{csxt} that are used as an indicator of data quality in that subset of the genealogical dataset.

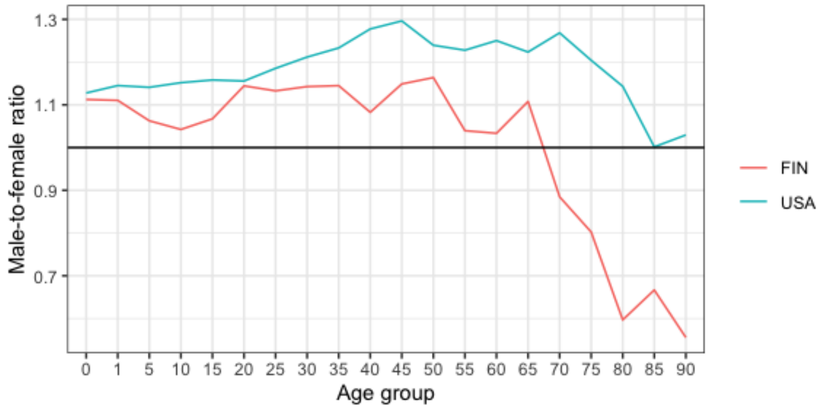
Specifically, we include:

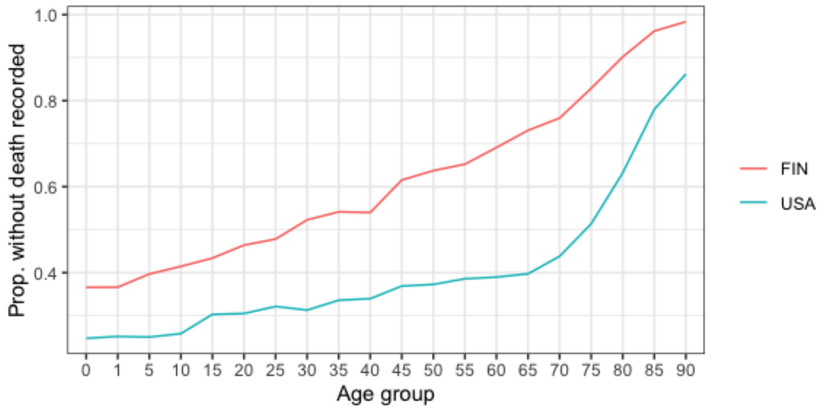
1. the (log-transformed) gender ratio

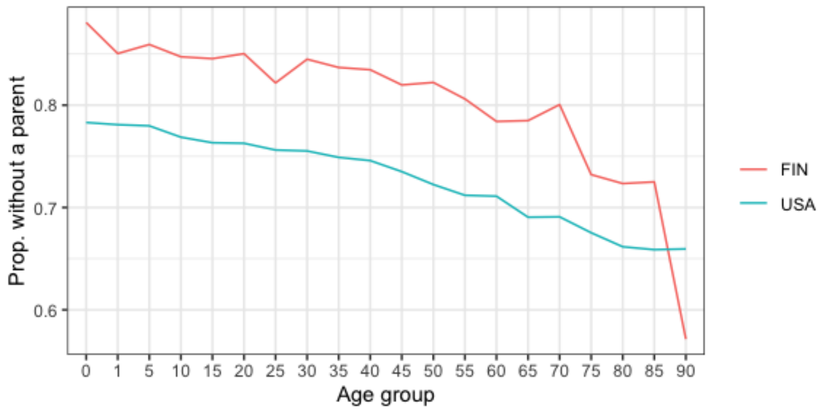
2. the (logit-transformed) proportion of records with no death date recorded

3. the (logit-transformed) proportion of records without a parent recorded

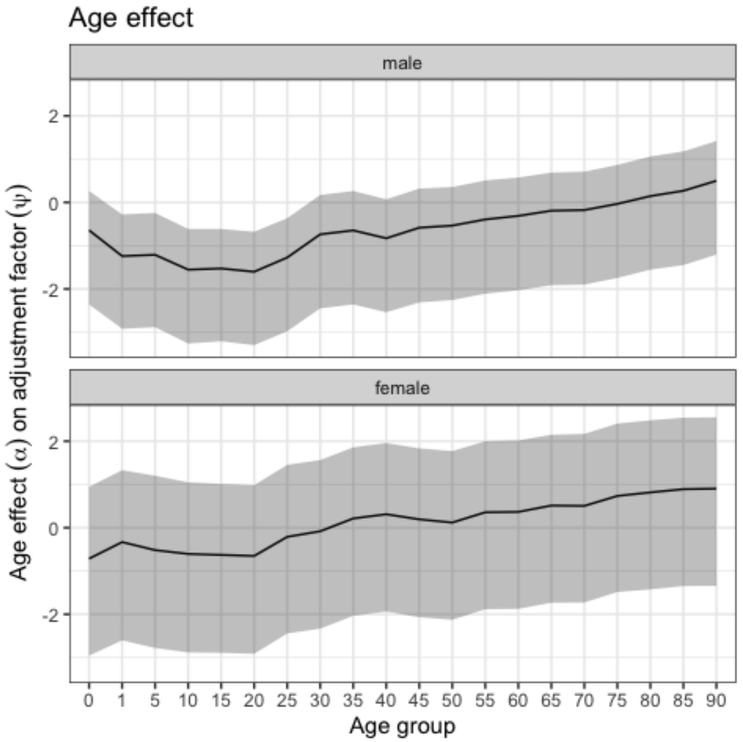
The goal of including these indicators is to allow for generalization to countries where we don't have high-quality data.



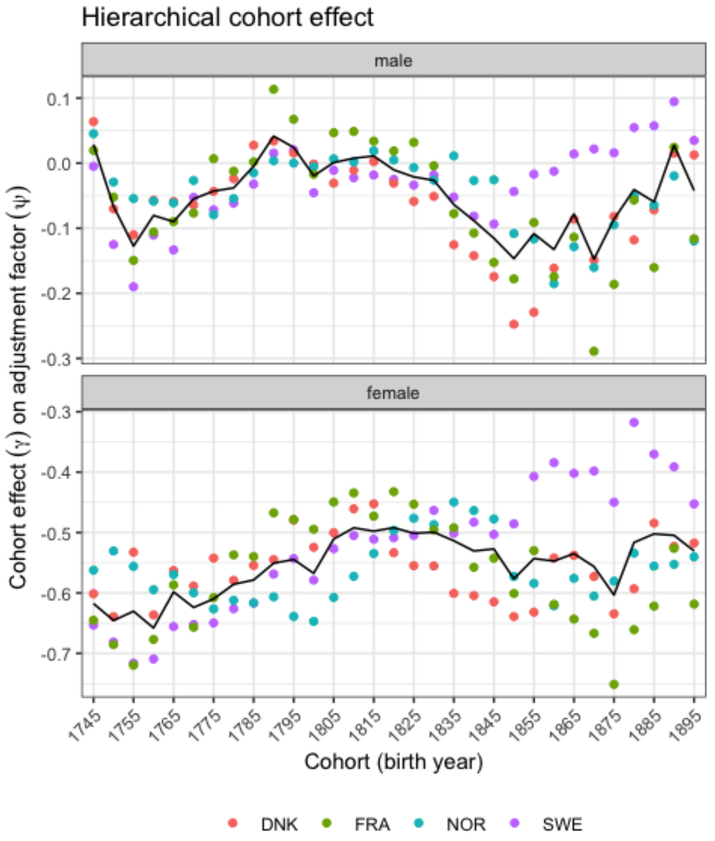




Estimated parameters

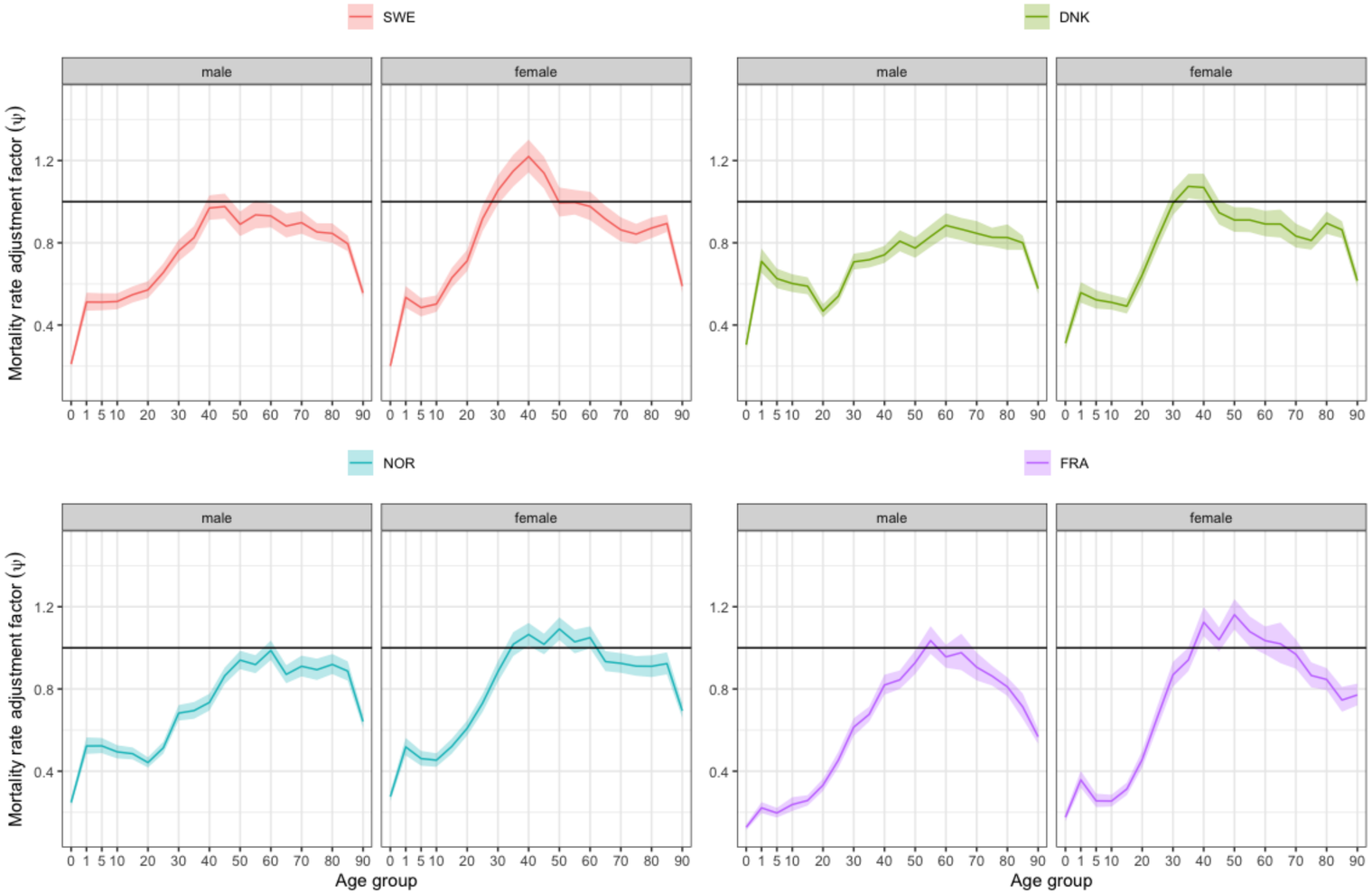


The adjustment factor ψ is allowed to vary over age. Based on the estimated age effects, Familinx mortality rates are relatively lower for younger ages, and increases for older ages.



A hierarchical cohort effect is also estimated for the adjustment factor. The mean cohort effect is shown with a solid black line, while points represent country-specific estimates.

Estimated adjustment factors

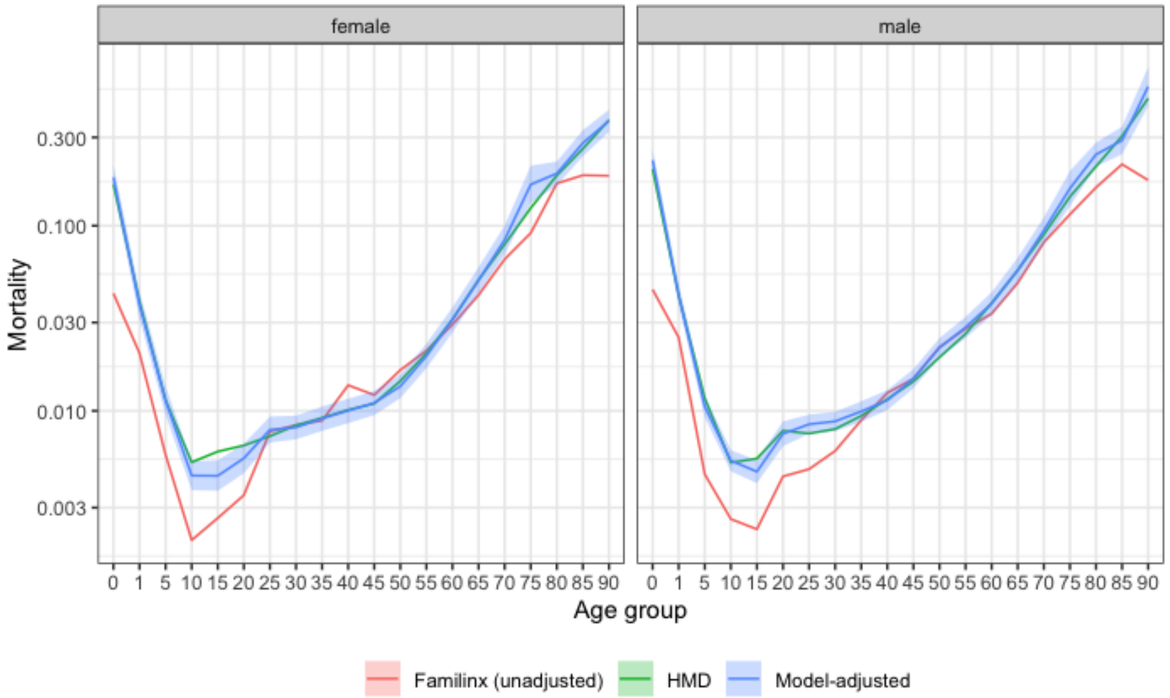


Modelled estimates of the **mortality rate adjustment factor ψ, representing the ratio of the Familinx mortality rate to the HMD rate,** for 1895 Sweden, Denmark, Norway, and France. Values less than 1 indicate that the Familinx mortality rates are too low, while values greater than 1 indicate that rates are too high.

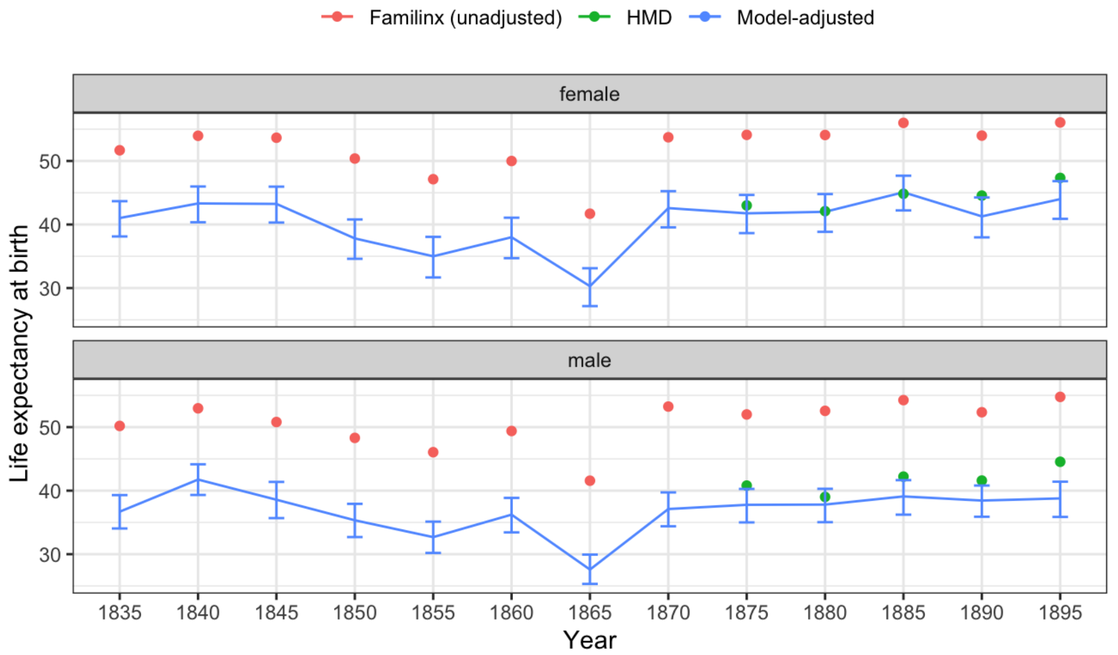
Inferred mortality from Familinx is usually too low, especially for young ages. Rates for middle-age groups (~40-65) are somewhat more representative, particularly for female populations, sometimes exceeding the HMD rate.

The representativeness of the Familinx-derived mortality rates varies between populations. Countries exhibit different age structures, degrees of bias, and occasionally directions of bias.

Finland as a test case

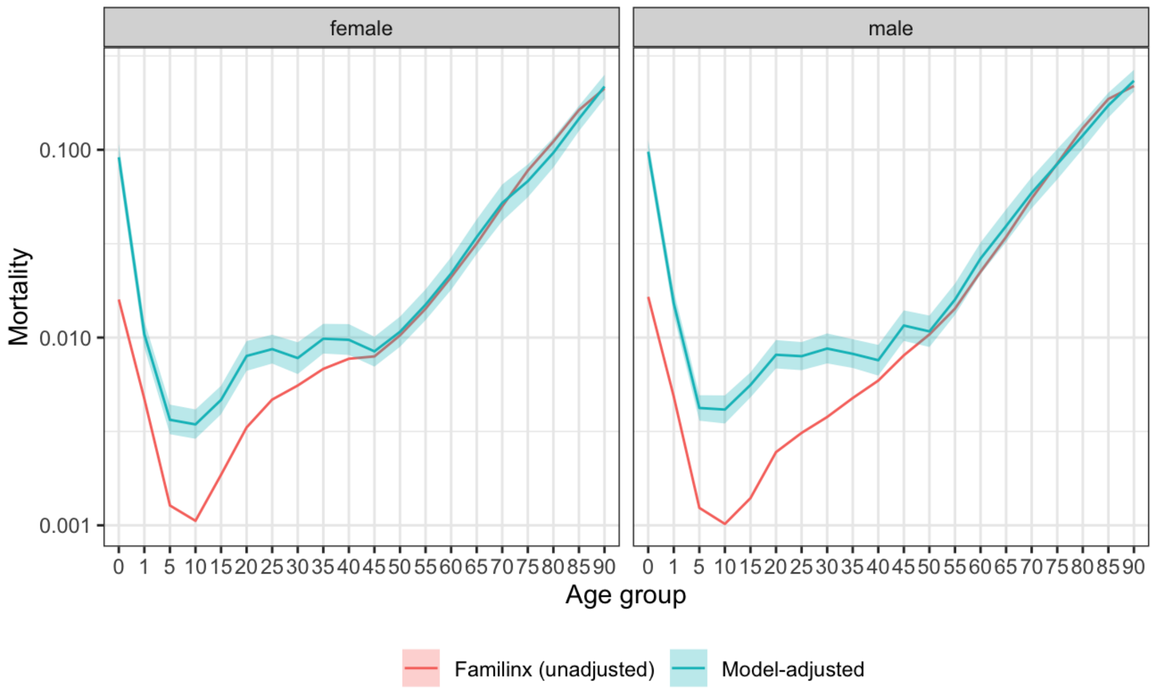


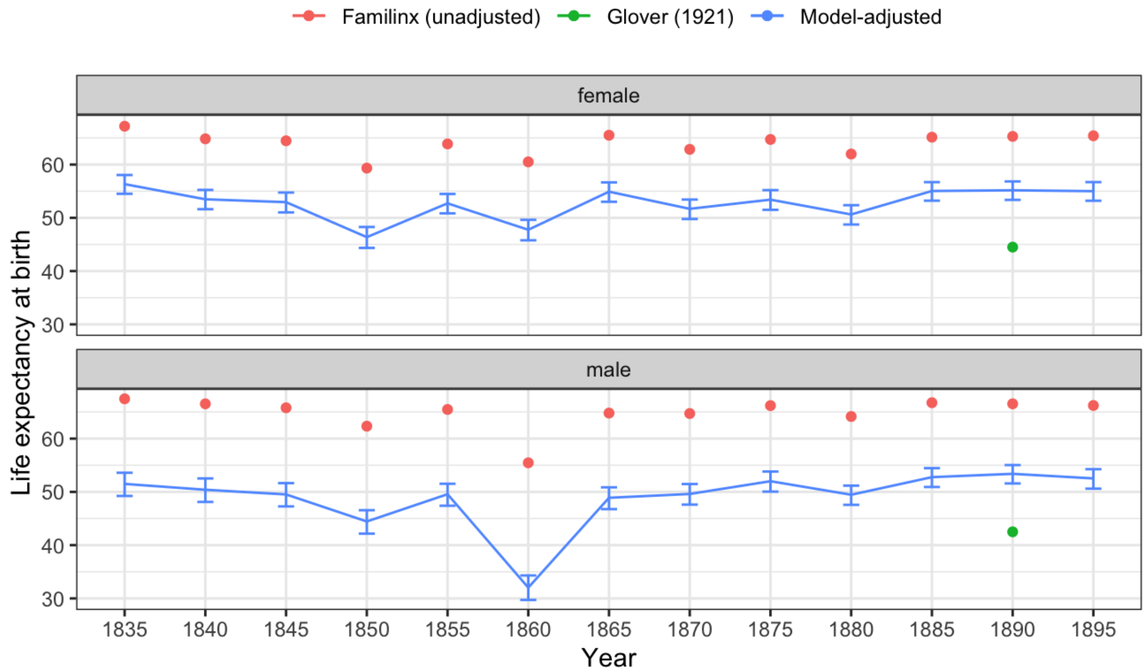
Mortality rates for 1880 Finland. Curves are coloured according to the data source, and the shaded blue region depicts a 80% credible interval for the model-adjusted estimate. **The model smooths the Familinx mortality curve, and corrects the rates upward for most age groups, resulting in an estimate closer to the HMD rate.**



Estimated life expectancies at birth for Finland by data source. Estimates from HMD rates are shown where available. Model-adjusted life expectancies are lower than the unadjusted data and are closer to the HMD values, but overcorrect in later years. The lower life expectancy 1865-69 may be explained by the Great Finnish Famine of 1866-68.

Application to the United States





Mortality rates for 1890 United States. Colours distinguish adjusted and unadjusted Familinx estimates, and the shaded blue region depicts a 80% credible interval for the model-adjusted estimate. **Mortality rates are adjusted upward for younger age groups, but are relatively unchanged for older age groups** (>50).

Estimated life expectancies at birth for the United States. High quality data for this period are not available, but an estimate for 1890 Massachusetts from a historical life table is shown in green [4]. **While the model predicts life expectancies to be much lower than the unadjusted Familinx estimates, they may still be too high given this reference point, indicating the need for further adjustment.**

Limitations and future work

Data extraction and imputation

In the data extraction process, we make simplifying assumptions about migration of individuals, which in some cases involve imputing the place of residence or the place of death. **Accounting for imputations in the statistical model** may better explain variation and more accurately reflect uncertainty, particularly for places with high migration.

Data quality indicators

The use of indicators in our model roughly capture some notion of data quality, and consequently misrepresentation in the demographic rates. However, **future work may focus on incorporating these measures more directly to explicitly model the data-generating process.** For example, knowledge that some proportion of deaths in the genealogy are unrecorded could be used to explicitly adjust death counts.

Generalization

Since our method relies on calibrating against a high quality data source, generalization to time periods or regions where high quality data are not available requires assumptions about how the genealogical data and the true demographic rates are related. While our test case results are encouraging, **it's not yet clear that the mortality rate biases are similar for all countries.**

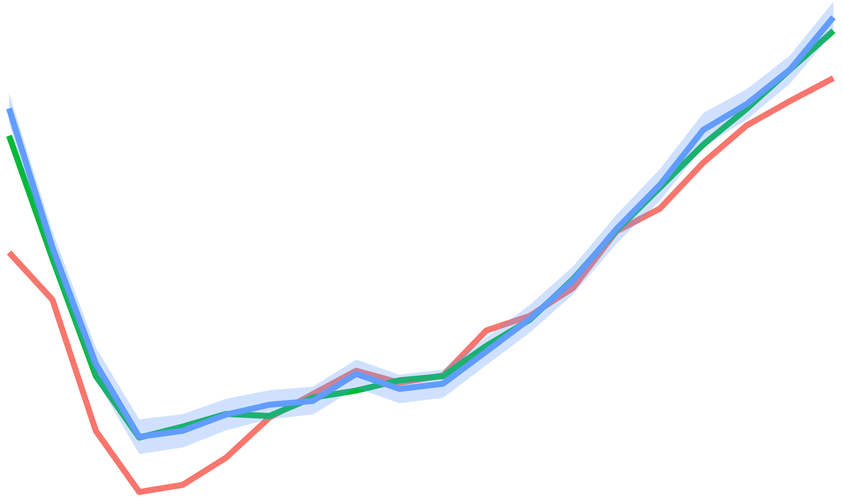
In particular, estimates are undercorrected if specific high-mortality groups are underrepresented in the genealogical data of particular country, but such groups do not exist in the countries used for calibration. In the United States test case for example, **the inability to account for marginalized populations who do not have access to recorded family histories may explain the over-optimistic estimates of life expectancy.**

Summary

In this study, we show that demographic rates extracted from large, crowdsourced genealogies can be severely biased. Implied mortality rates are usually too low, particularly for younger ages.

We demonstrate that this bias can be mitigated using a statistical model that calibrates the rates against a high quality data source, and apply our procedure to produce estimates of mortality and life expectancy for 1800s Finland and United States.

While online genealogical datasets hold potential for novel demographic inferences, researchers should exercise caution in drawing conclusions that may be subject to unrepresentativeness in these data.



References

[1] 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.

[2] 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.

[3] *Human Mortality Database.* University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). Available at www.mortality.org or www.humanmortality.de (data downloaded on [date])

[4] 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.