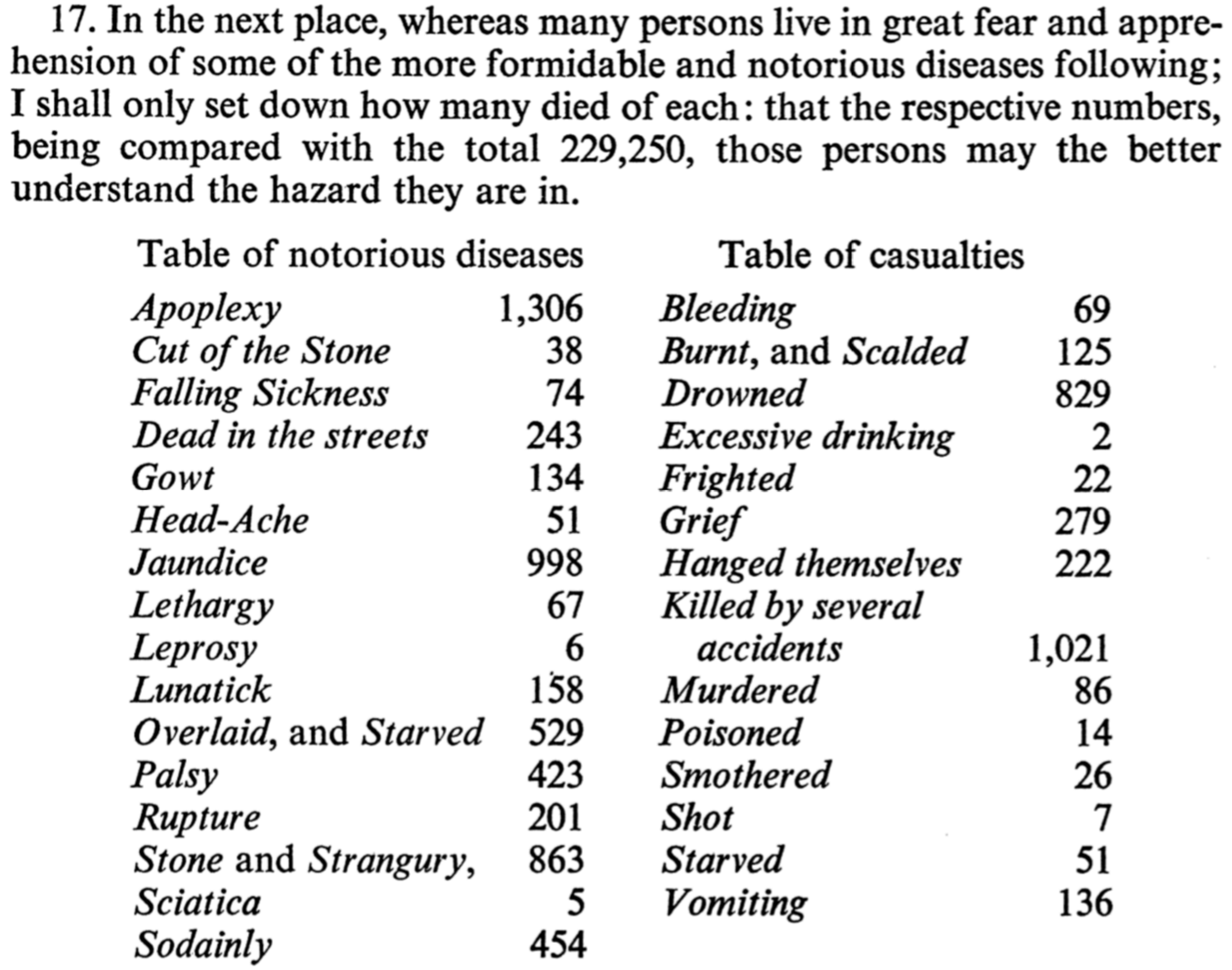
Estimating historical mortality rates using crowd-sourced online genealogies

Monica Alexander, Statistical Sciences and Sociology, University of Toronto

Background

- Historical mortality is important to understand
  - Relationship between industrialization/ urbanization and health
  - Differences by geography, social class
- But data are limited, particularly before 1900
  - Civil registration systems generally did not exist
  - Rely on parish records, city records, approximations



17. In the next place, whereas many persons live in great fear and apprehension of some of the more formidable and notorious diseases following; I shall only set down how many died of each: that the respective numbers, being compared with the total 229,250, those persons may the better understand the hazard they are in.

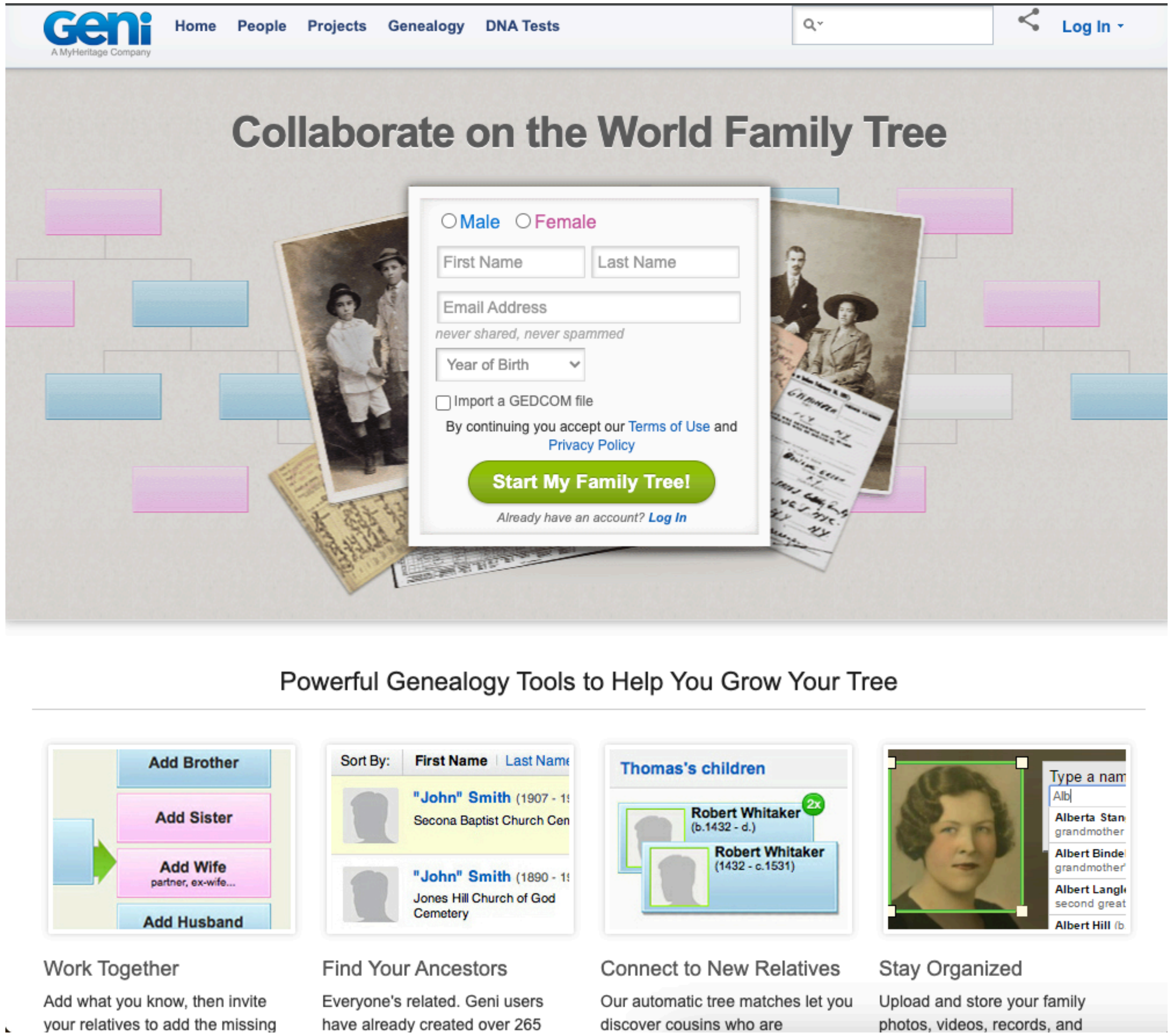
Table of notorious diseases	
Apoplexy	1,306
Cut of the Stone	38
Falling Sickness	74
Dead in the streets	243
Gowt	134
Head-Ache	51
Jaundice	998
Lethargy	67
Leprosy	6
Lunatick	158
Overlaid, and *Starved*	529
Palsy	423
Rupture	201
Stone and *Strangury*,	863
Sciatica	5
Sodainly	454

Table of casualties	
Bleeding	69
Burnt, and *Scalded*	125
Drowned	829
Excessive drinking	2
Frighted	22
Grief	279
Hanged themselves	222
Killed by several accidents	1,021
Murdered	86
Poisoned	14
Smothered	26
Shot	7
Starved	51
Vomiting	136

From Grant (1667)

The emergence of online genealogies

- User-entered information on family trees
- Ascendant genealogies, ancestry is reconstructed retrospectively
- Social networking aspect



Collaborate on the World Family Tree

○ Male ○ Female

First Name | Last Name

Email Address

never shared, never spammed

Year of Birth

□ Import a GEDCOM file

By continuing you accept our Terms of Use and Privacy Policy

Start My Family Tree!

Already have an account? Log In

Powerful Genealogy Tools to Help You Grow Your Tree

"John" Smith (1907 - 1!
Secona Baptist Church Cen

"John" Smith (1890 - 1!
Jones Hill Church of God Cemetery

Robert Whitaker (b.1432 - d.)

Robert Whitaker (1432 - c.1531)

Alberta Stan
grandmother

Albert Binde
grandmother

Albert Langl
second great

Albert Hill (b

Work Together

Add what you know, then invite your relatives to add the missing

Find Your Ancestors

Everyone's related. Geni users have already created over 265

Connect to New Relatives

Our automatic tree matches let you discover cousins who are

Stay Organized

Upload and store your family photos, videos, records, and

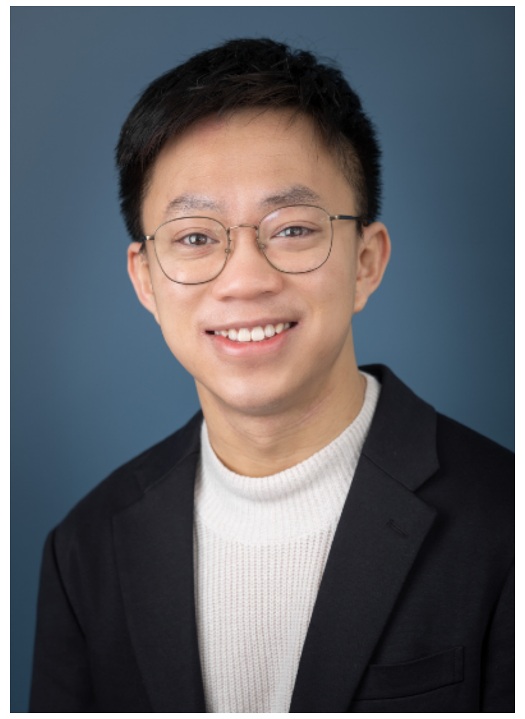
Online genealogies as a demographic data source

- **Familinx**: 86 million individual records
- Demographic events: births, deaths, marriages
- Kinship ties: (known for 43 million individuals)
- Mostly in the Global North (85% of vital events occur in Europe or North America)
- Family ties not restricted by geographic boundaries
- But some clear issues (ascendants, selective remembering, selection of users, misreporting)

This project

- Goal: Construct historical mortality rates and life expectancies for a group of countries between 1800-1900
- Developed a Bayesian model to estimate and correct for biases in mortality rates derived from Familinx
- Working paper: https://www.demogr.mpg.de/papers/working/wp-2022-005.pdf
- Code etc: https://github.com/michael-chong/familinx-mort

Collaborators:



Michael Chong
Statistics
University of Toronto

Diego Alburez
Kinequalities Group
Max Planck Institute for
Demographic Research

Emilio Zagheni
Digital Demography Group
Max Planck Institute for
Demographic Research

Data

Familinx overview

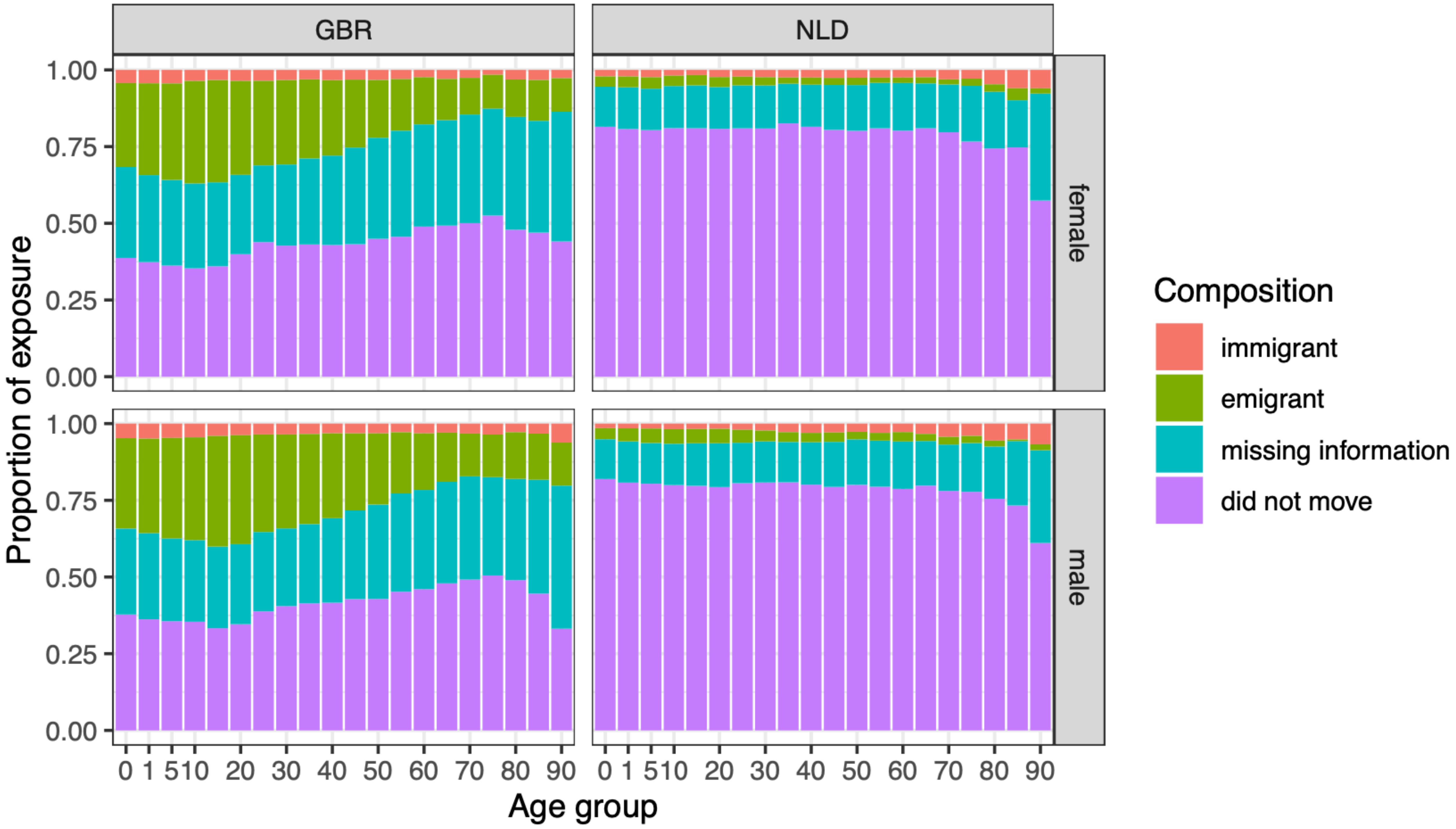
- Consider ten countries in the period 1800-1900: Belgium, Denmark, Finland, France, Netherlands, Norway, Sweden, Switzerland, UK, USA
- ~5.5 million individual records
- We tabulate deaths and exposure (person-years) by age group, time period, gender, and country
- Data issues:
  - Geographic names in free text (errors, non-English, historical names)
  - Some locations of birth/death missing
  - No direct information on the place of residence over the life course

Constructing exposure counts from Familinx

- Exposures to death are divided among the countries in which they had vital events (birth, birth of child, death)

- Sometimes successive locations are missing

- In general, migrants contribute a small share of the exposure to death counts (although the UK is an exception)



Gold-standard data

- Human Mortality Database: high-quality repository of harmonized mortality data

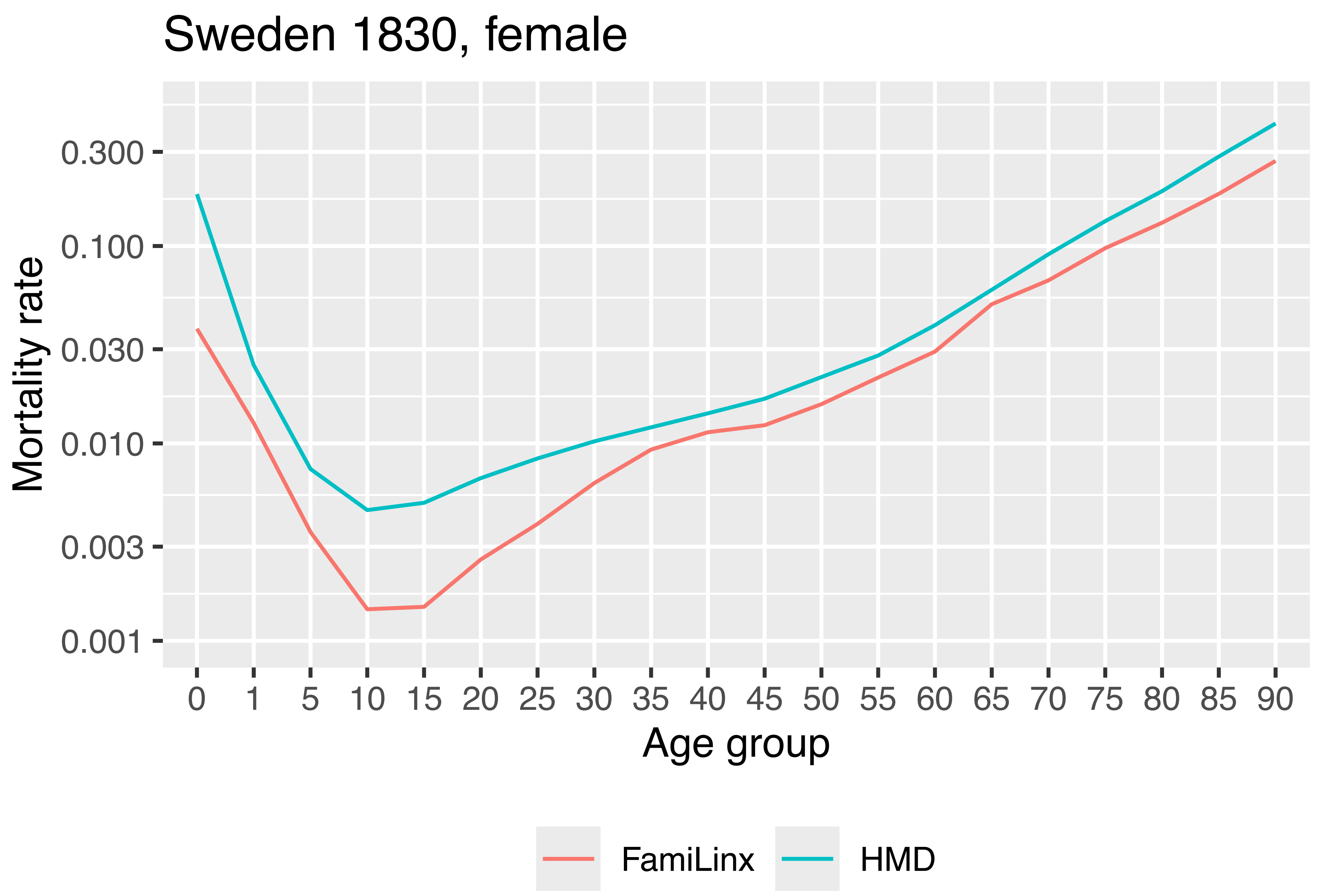
- Obtain death counts and exposure (person-years lived) by country, gender, age, and time period

Country	**Earliest year available in Human Mortality Database**
Sweden	1751
France	1816
Denmark	1835
Belgium	1841
Great Britain	1841
Norway	1846
Netherlands	1850
Switzerland	1876
Finland	1878

Model

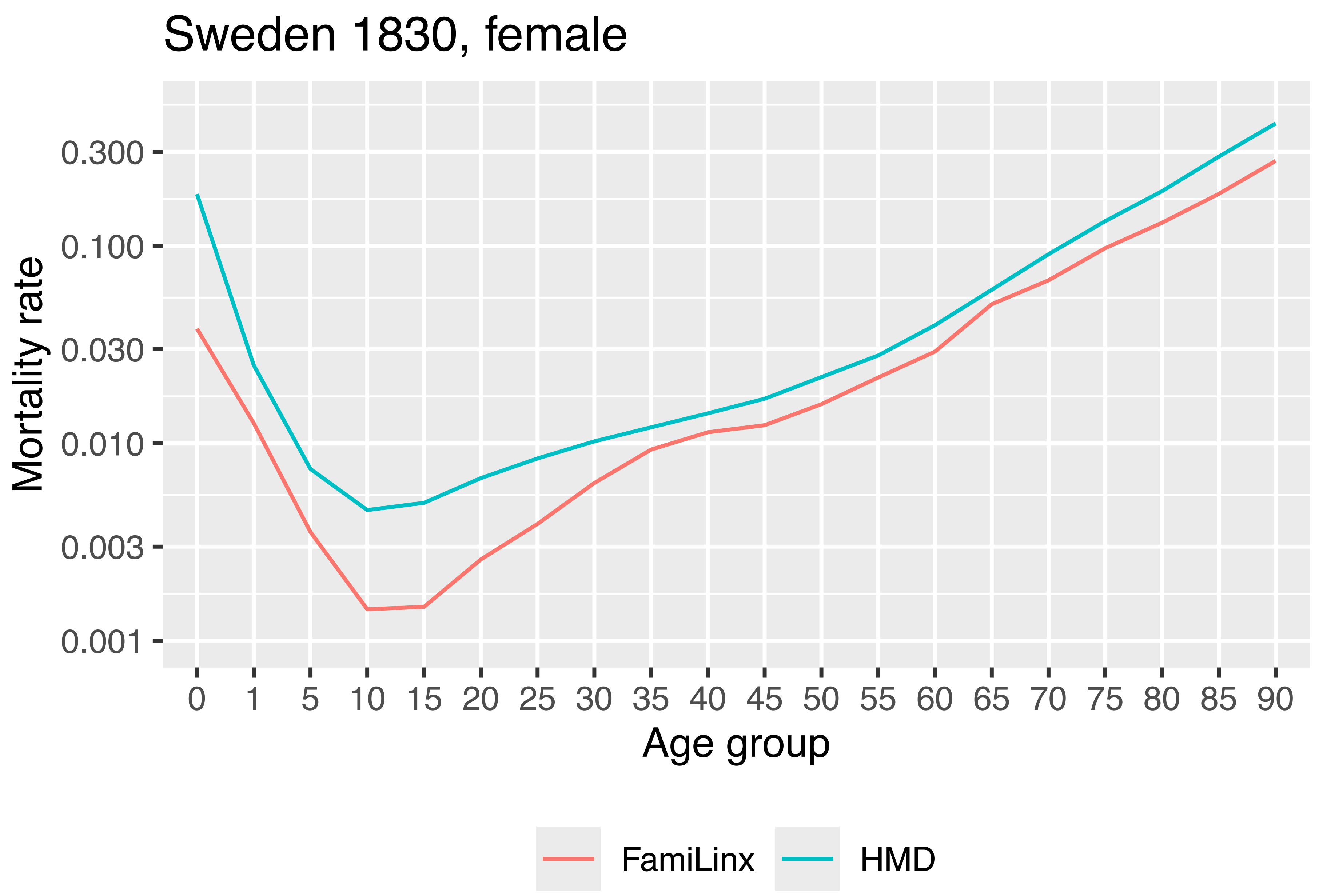
Modeling goals

In some country-years we have both Familinx and HMD data



Modeling goals

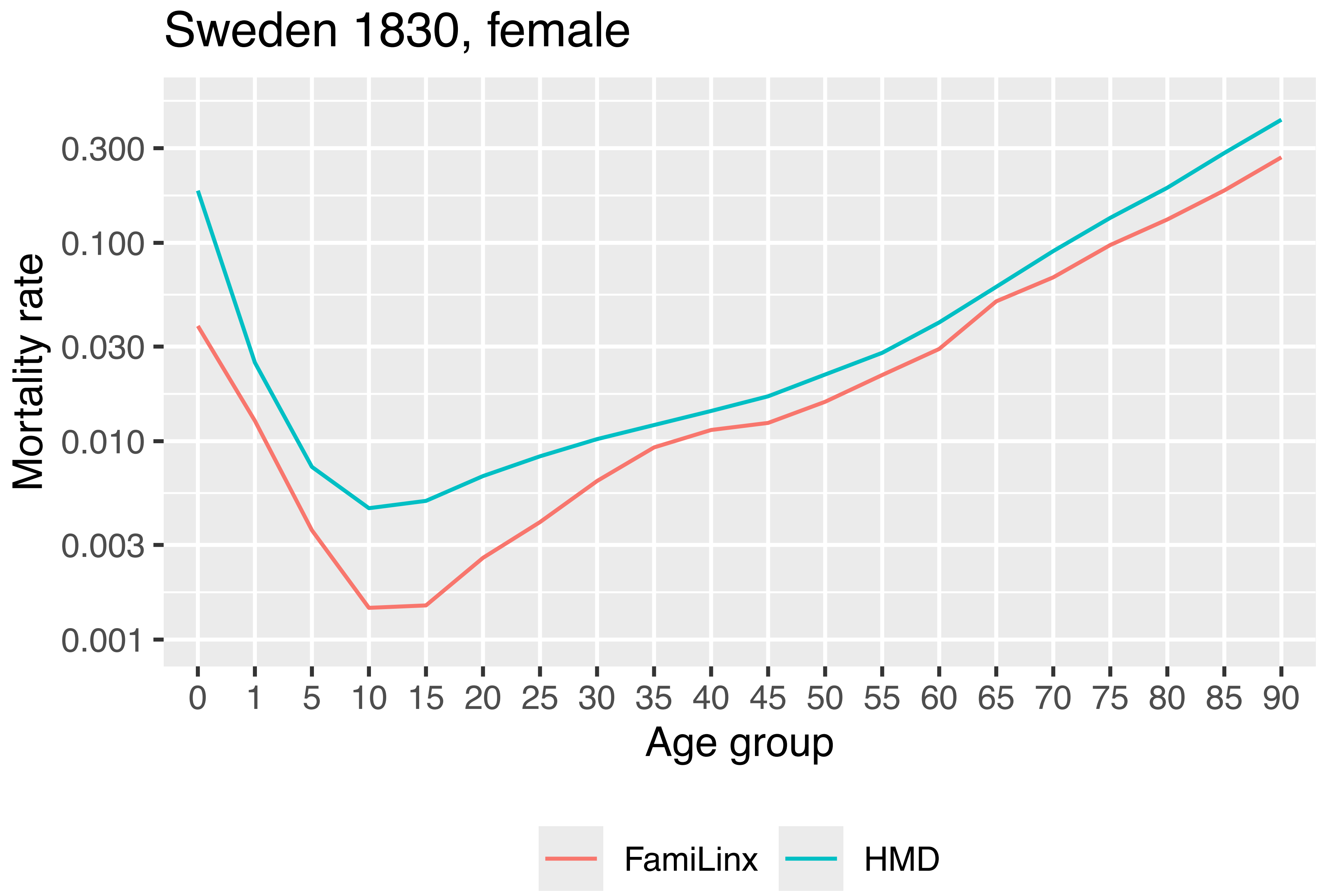
In some country-years we have both Familinx and HMD data

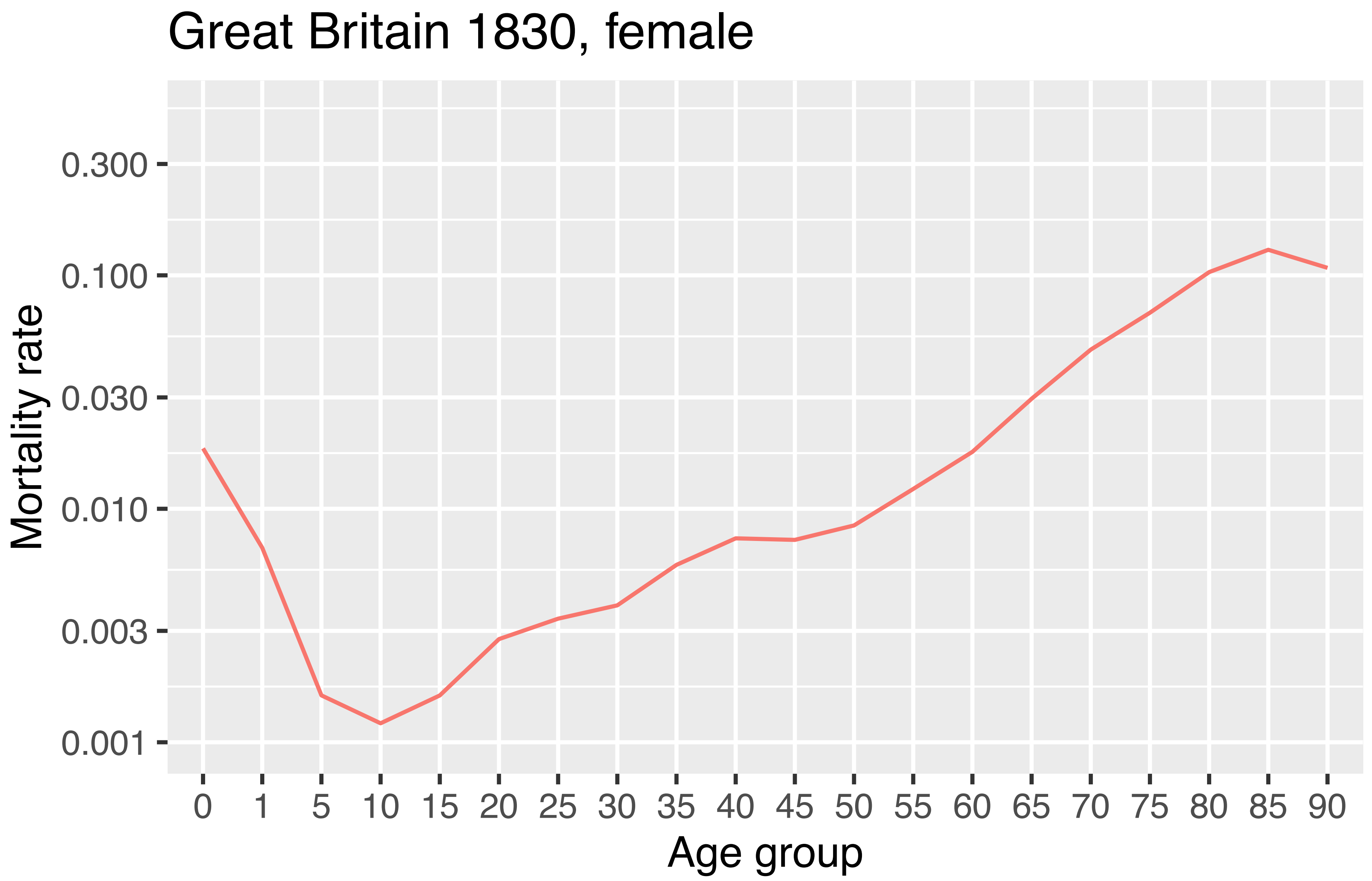


In these cases, could directly calculate a set of ‘adjustment factors’ by age for a particular country / year/ gender to adjust the familinx mortality data

Modeling goals

But for many country-years we only have Familinx

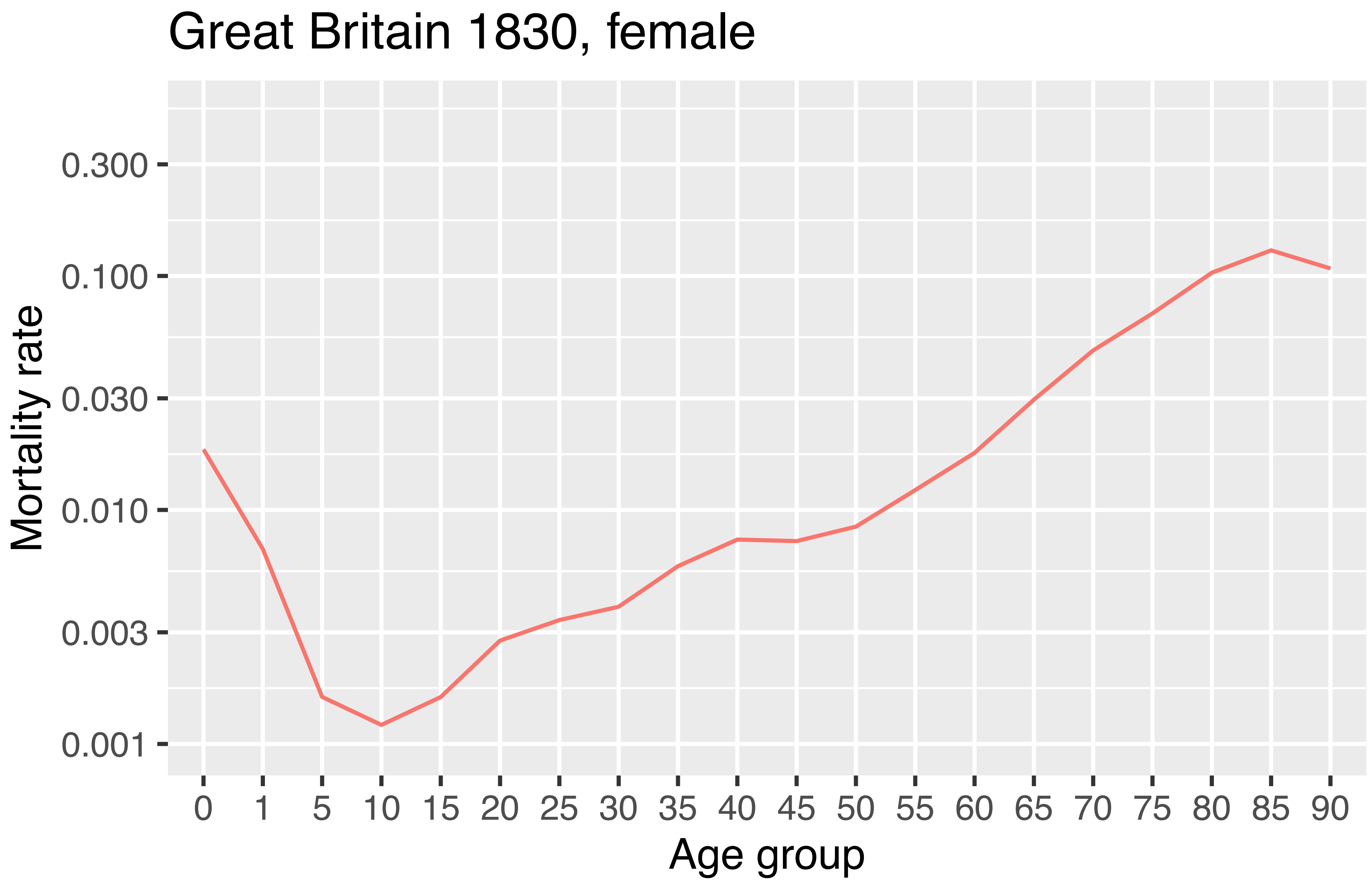




Modeling goals

But for many country-years we only have Familinx

So need a model to estimate a set of adjustment factors



Modeling goals

- Estimate age-specific mortality rates for each country and five-year periods between 1800 and 1900
- Account for
    - Characteristic shape of mortality over age
    - Biases in Familinx data share similarities across age/years/countries
    - Mortality rates tend to evolve smoothly over age and time (although sometimes don’t)

Notation

- Latent mortality rate for country *c*, gender *g*, time period *t,* age group *x:* $\mu_{c,g,t,x}$
- Observed death and exposure counts from HMD: $d^{(H)}_{c,g,t,x}, P^{(H)}_{c,g,t,x}$
- Observed death and exposure counts from Familinx: $d^{(F)}_{c,g,t,x}, P^{(F)}_{c,g,t,x}$
- Adjustment factors for Familinx: $\psi_{c,g,t,x}$

Data models

- For HMD

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

- For Familinx

$$d^{(F)}_{c,g,t,x} \,|\, \mu_{c,g,t,x}, \phi^{(F)}_x \sim \text{NegBinom}\left(\mu_{c,g,t,x} \psi_{c,g,t,x} P^{(F)}_{c,g,t,x}, \phi^{(F)}_x\right)$$

Data models

- For HMD

$$d^{(H)}_{c,g,t,x} \,|\, \mu_{c,g,t,x}, \phi^{(H)}_x \sim \text{NegBinom}\left(\boxed{\mu_{c,g,t,x}} P^{(H)}_{c,g,t,x}, \phi^{(H)}_x\right)$$

- For Familinx

$$d^{(F)}_{c,g,t,x} \,|\, \mu_{c,g,t,x}, \phi^{(F)}_x \sim \text{NegBinom}\left(\boxed{\mu_{c,g,t,x}}\boxed{\psi_{c,g,t,x}} P^{(F)}_{c,g,t,x}, \phi^{(F)}_x\right)$$

Mortality model

Mortality model (for $\mu_{c,g,t,x}$)

$$d^{(F)}_{c,g,t,x} \mid \mu_{c,g,t,x}, \phi^{(F)}_x \sim \text{NegBinom}\left(\boxed{\mu_{c,g,t,x}}\psi_{c,g,t,x} P^{(F)}_{c,g,t,x}, \phi^{(F)}_x\right)$$

- Mortality rates over age are expected to conform to characteristic shapes (‘J’ shape)

- We incorporate this assumption by modelling the logged mortality rates as a linear combination of ‘principal components’, which are derived from gold-standard data (HMD)

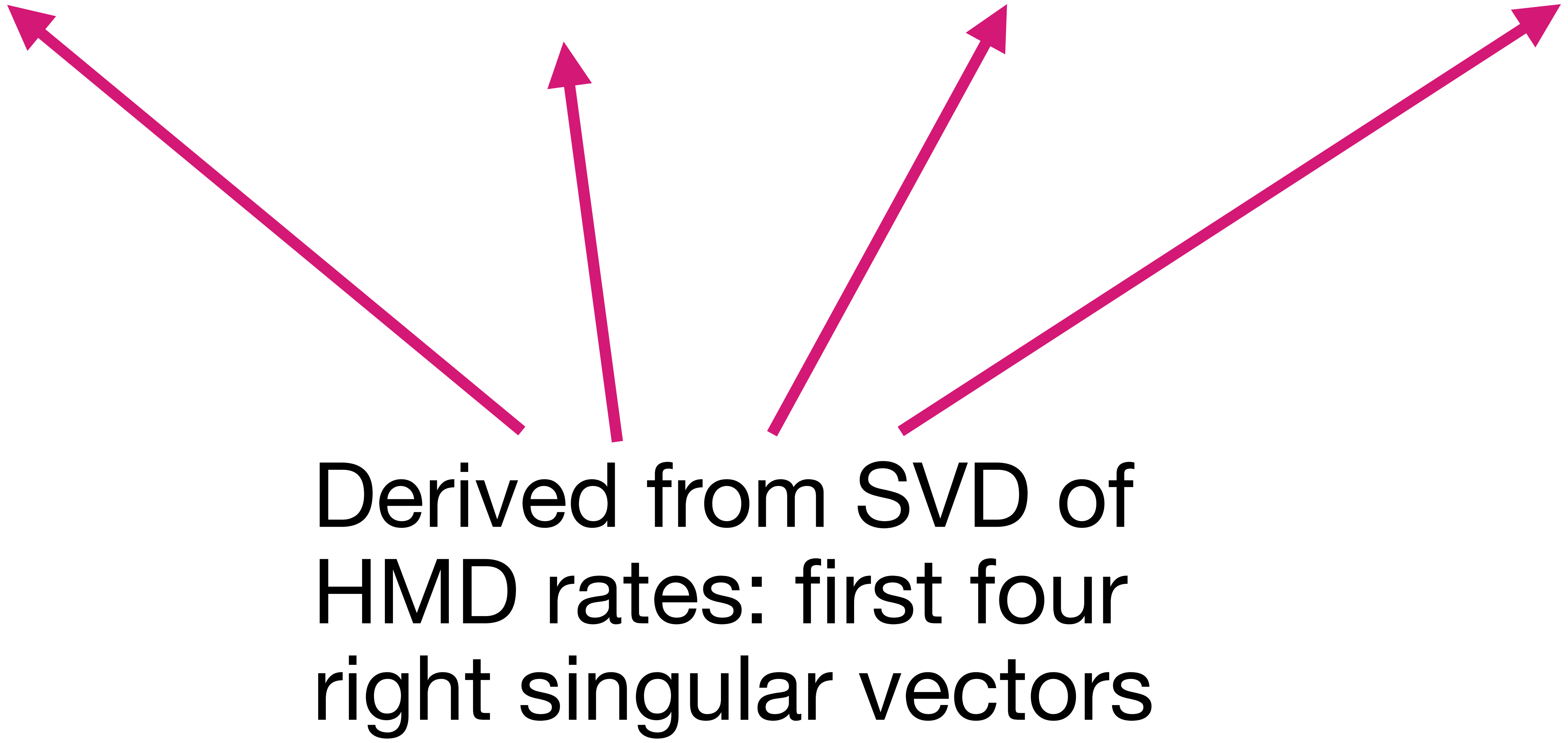
- We also allow for historical mortality shocks by including a regularized error term

Mortality model

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

Mortality model

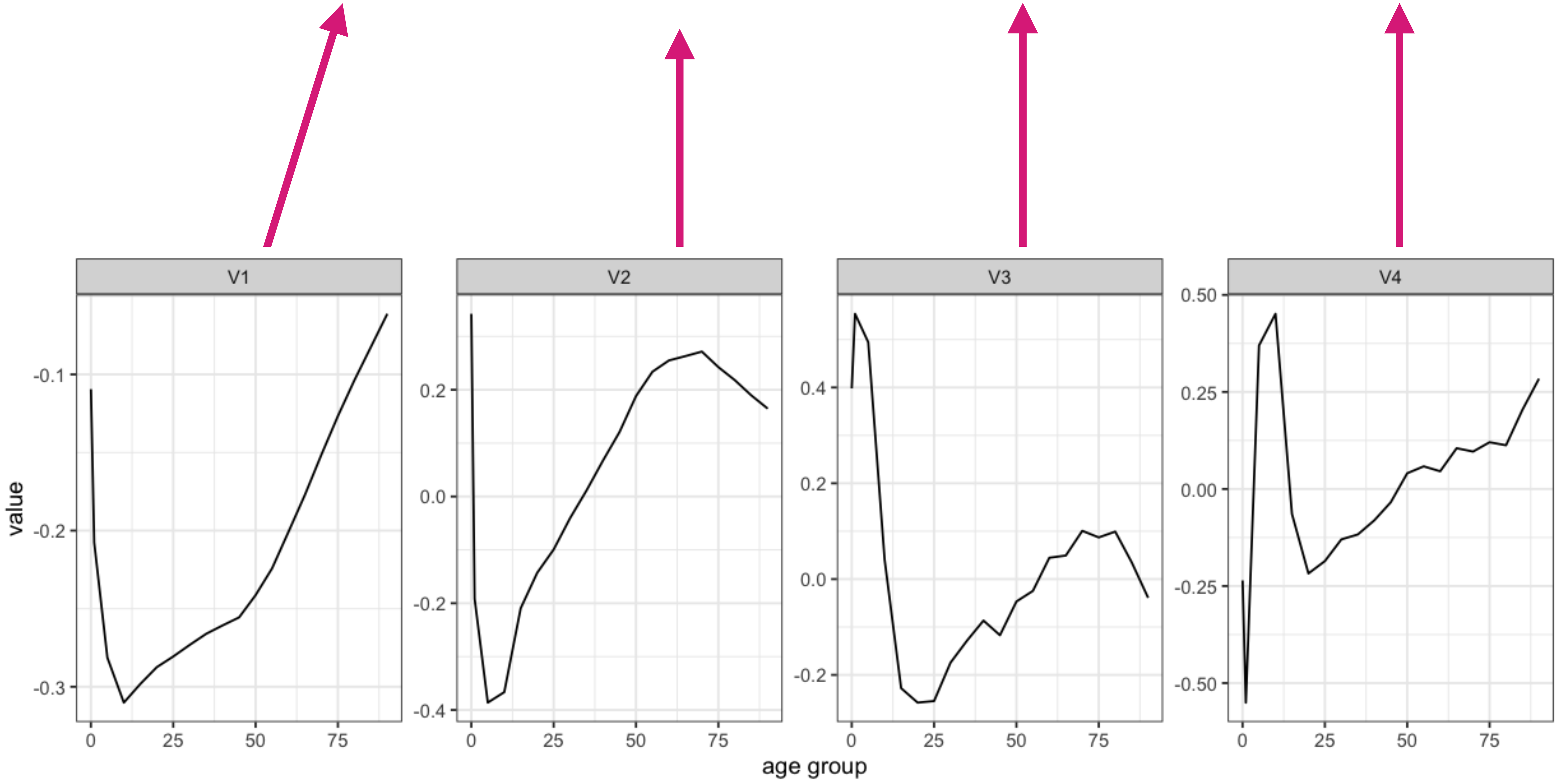
$$\log \overrightarrow{\mu}_{c,g,t} = \eta_{1,c,g,t}\vec{v}_{g,1} + \eta_{2,c,g,t}\vec{v}_{g,2} + \eta_{3,c,g,t}\vec{v}_{g,3} + \eta_{4,c,g,t}\vec{v}_{g,4} + \vec{\varepsilon}_{c,g,t}$$



Derived from SVD of
HMD rates: first four
right singular vectors

Mortality model

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



Mortality model

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

Coefficients assumed to change smoothly over time

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

Cubic basis splines, second-order penalty

Mortality model

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

Error term, assume
mostly close to zero
unless detectable shock
(e.g. war)

Mortality model

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

Regularized horseshoe prior:

$$\varepsilon_i \mid \tau, \lambda_i \sim N\left(0,\tau^2\tilde{\lambda}_i^2\right)$$

$$\tilde{\lambda}_i^2 = \frac{d^2\lambda_i^2}{d^2 + \tau^2\lambda_i^2}$$

$$\tau \sim \text{Cauchy}\,(0,0.01)$$

$$\lambda_i^2 \sim \text{Cauchy}(0,1)$$

Model for adjustment factors

Adjustment factor

$$d^{(F)}_{c,g,t,x} \mid \mu_{c,g,t,x}, \phi^{(F)}_x \sim \text{NegBinom}\left(\mu_{c,g,t,x}\boxed{\psi_{c,g,t,x}}P^{(F)}_{c,g,t,x}, \phi^{(F)}_x\right)$$

Goal: flexibly capture patterns in Familinx bias across age, time, gender, country

Adjustment factor

$$d^{(F)}_{c,g,t,x} \,|\, \mu_{c,g,t,x}, \phi^{(F)}_x \sim \text{NegBinom}\left(\mu_{c,g,t,x}\boxed{\psi_{c,g,t,x}}P^{(F)}_{c,g,t,x}, \phi^{(F)}_x\right)$$

Goal: flexibly capture patterns in Familinx bias across age, time, gender, country

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

Adjustment factor

Goal: flexibly capture patterns in Familinx bias across age, time, gender, country

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

Country-specific intercept

Adjustment factor

Goal: flexibly capture patterns in Familinx bias across age, time, gender, country

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

Young age adjustments

Adjustment factor

Goal: flexibly capture patterns in Familinx bias across age, time, gender, country

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

2D tensor product smoothing spline

Adjustment factor

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

Has the form (for each *c* and *g*)

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

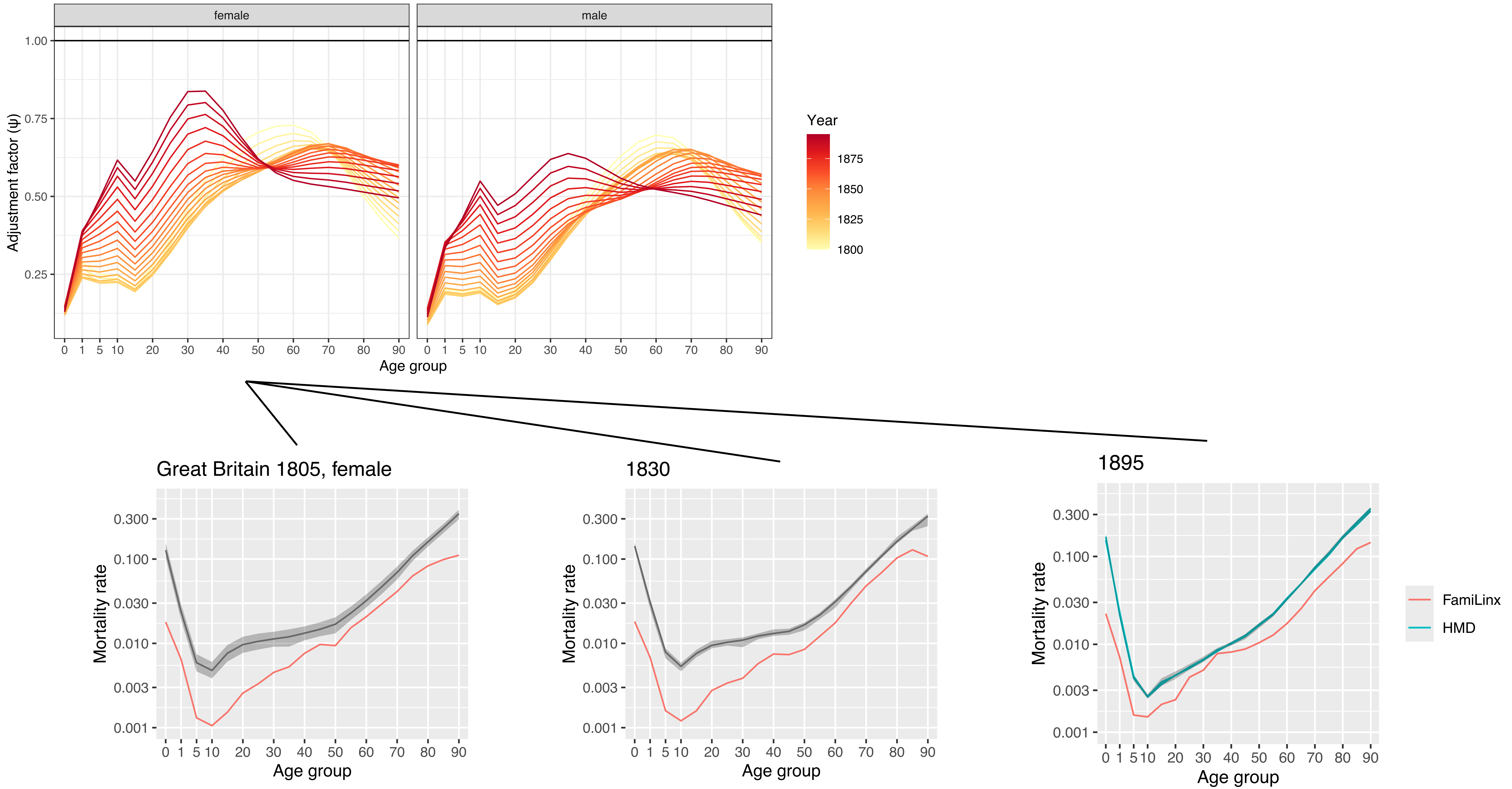
Coefficients (estimated hierarchically across countries)

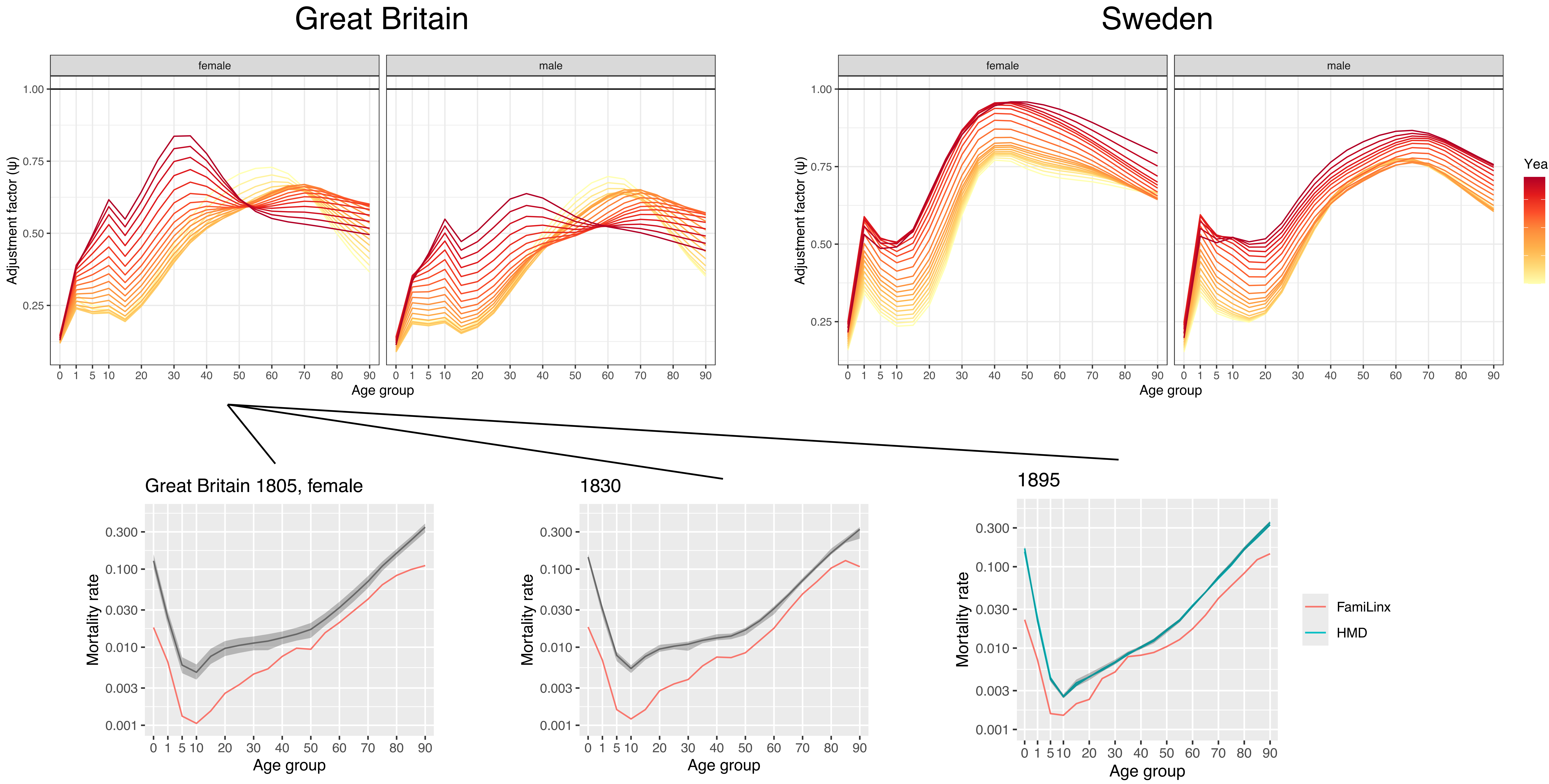
(second-order penalty on smoothness)

Cubic basis splines in
age and time dimension

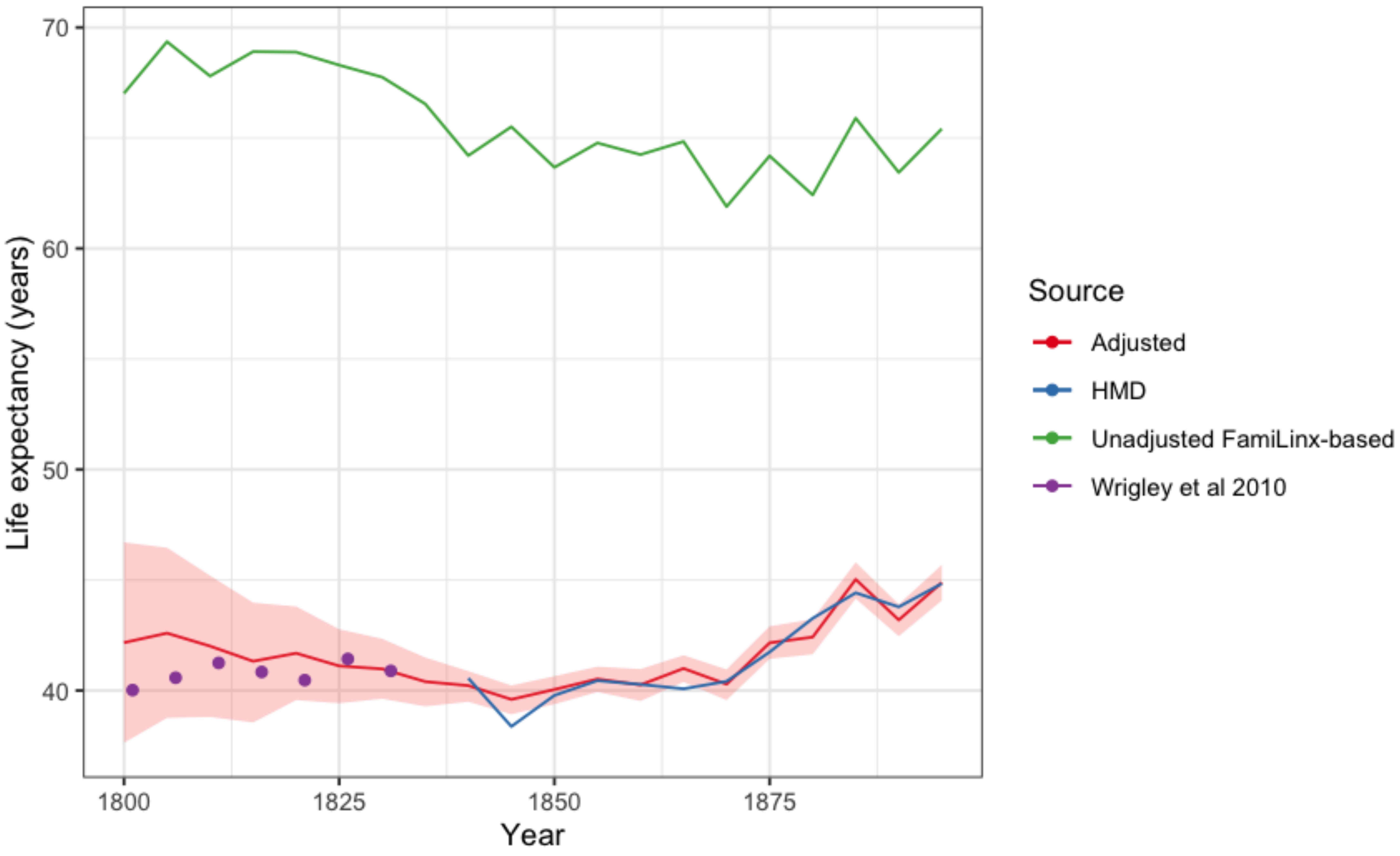
Illustrative Results

Great Britain

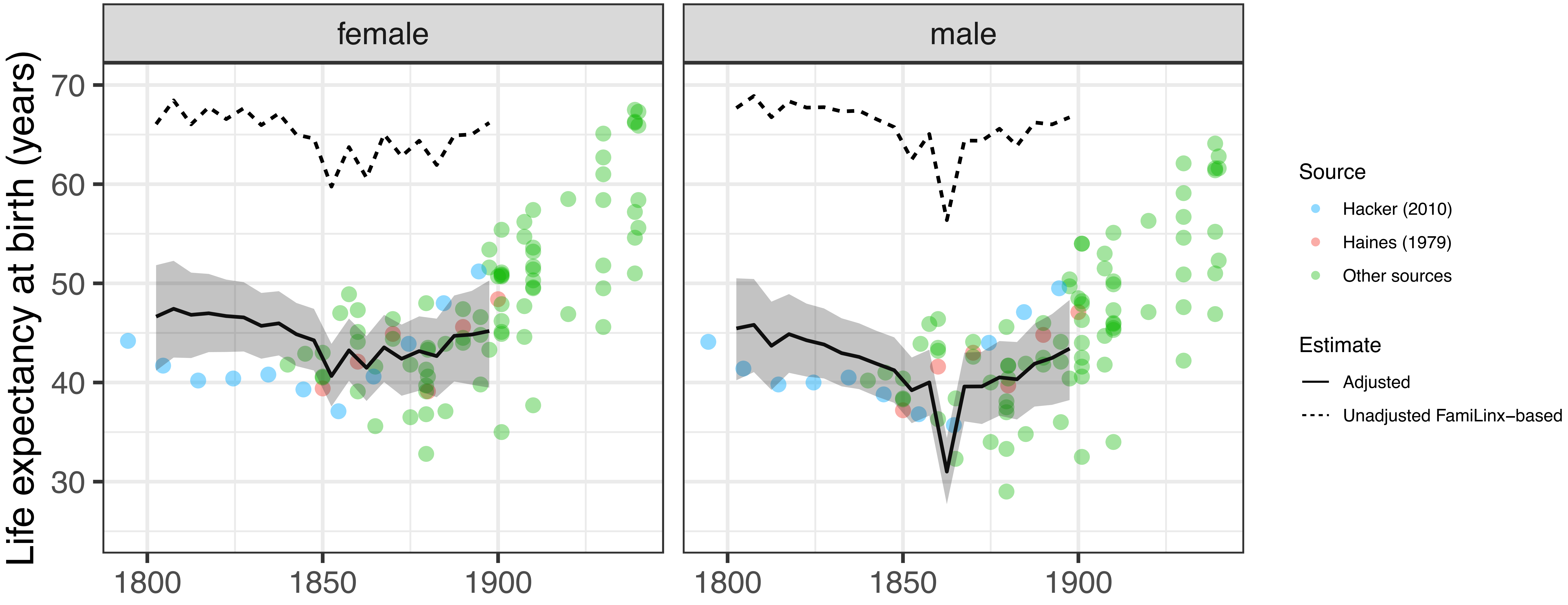



Great Britain
Sweden
female
male
female
male
1.00
0.75
0.50
0.25
Adjustment factor (ψ)
0 1 5 10 20 30 40 50 60 70 80 90
Age group
Year
1875
1850
1825
1800
Great Britain 1805, female
1830
1895
0.300
0.100
0.030
0.010
0.003
0.001
Mortality rate
Age group
FamiLinx
HMD

Life expectancy in Great Britain

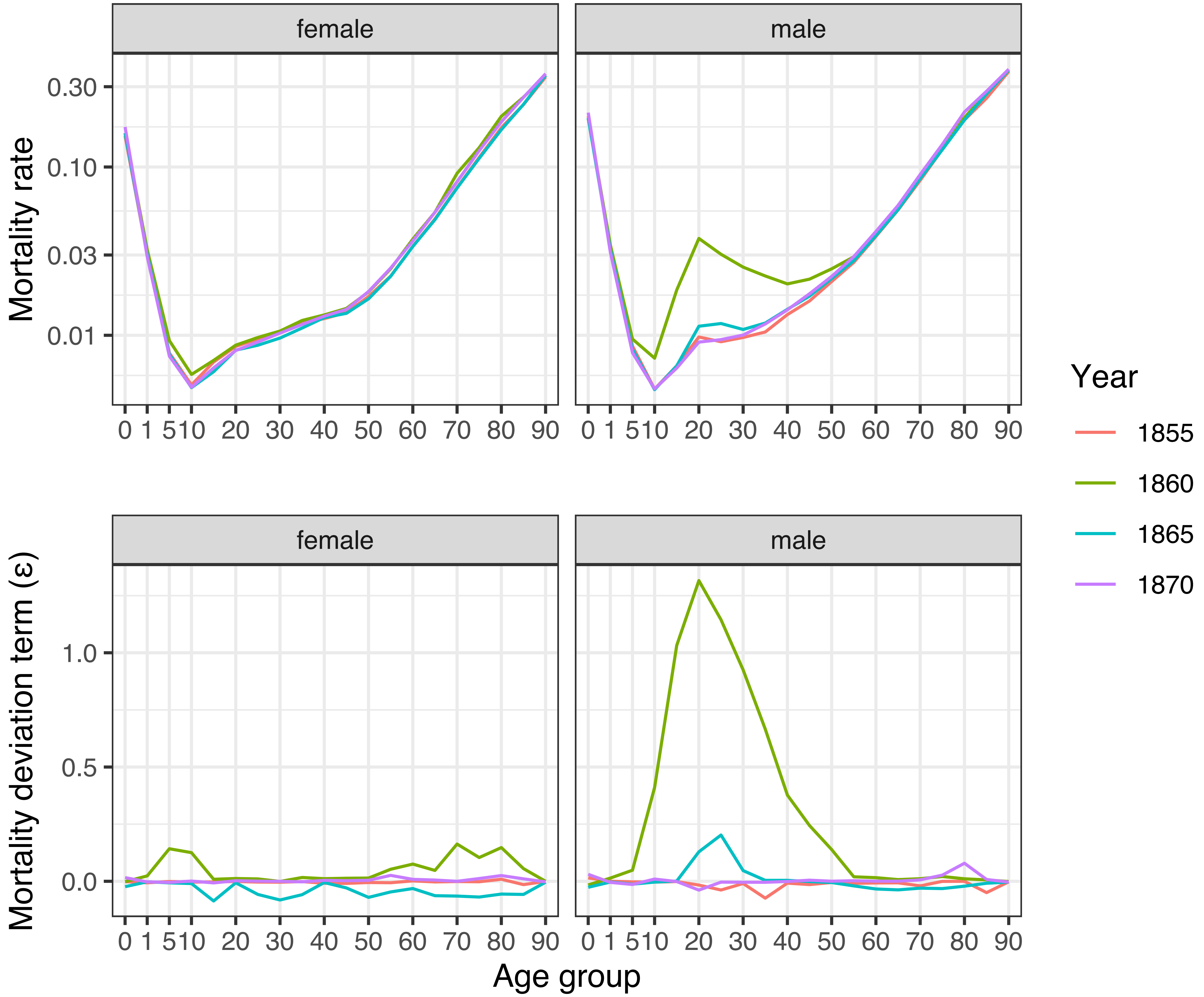


US life expectancy



Impact of regularized error term

Estimates for the United States in the Civil War years (1861-1865)



Systematic mortality rate variation is constrained to only vary in PC directions, and variation is smoothed over time

Leftover variation can be absorbed by the sparse ε terms

Summary

- Online genealogies offer a new potential source for historical demography
    - Not just mortality! fertility, kin structures
- Data are biased because of both how genealogies are reported but also because of quality, representativeness
- Bayesian model centered on estimating and predicting a set of adjustment factors, which are useful in their own right
- Future work will investigate generative models of genealogical construction

Thanks!

monica.alexander@utoronto.ca

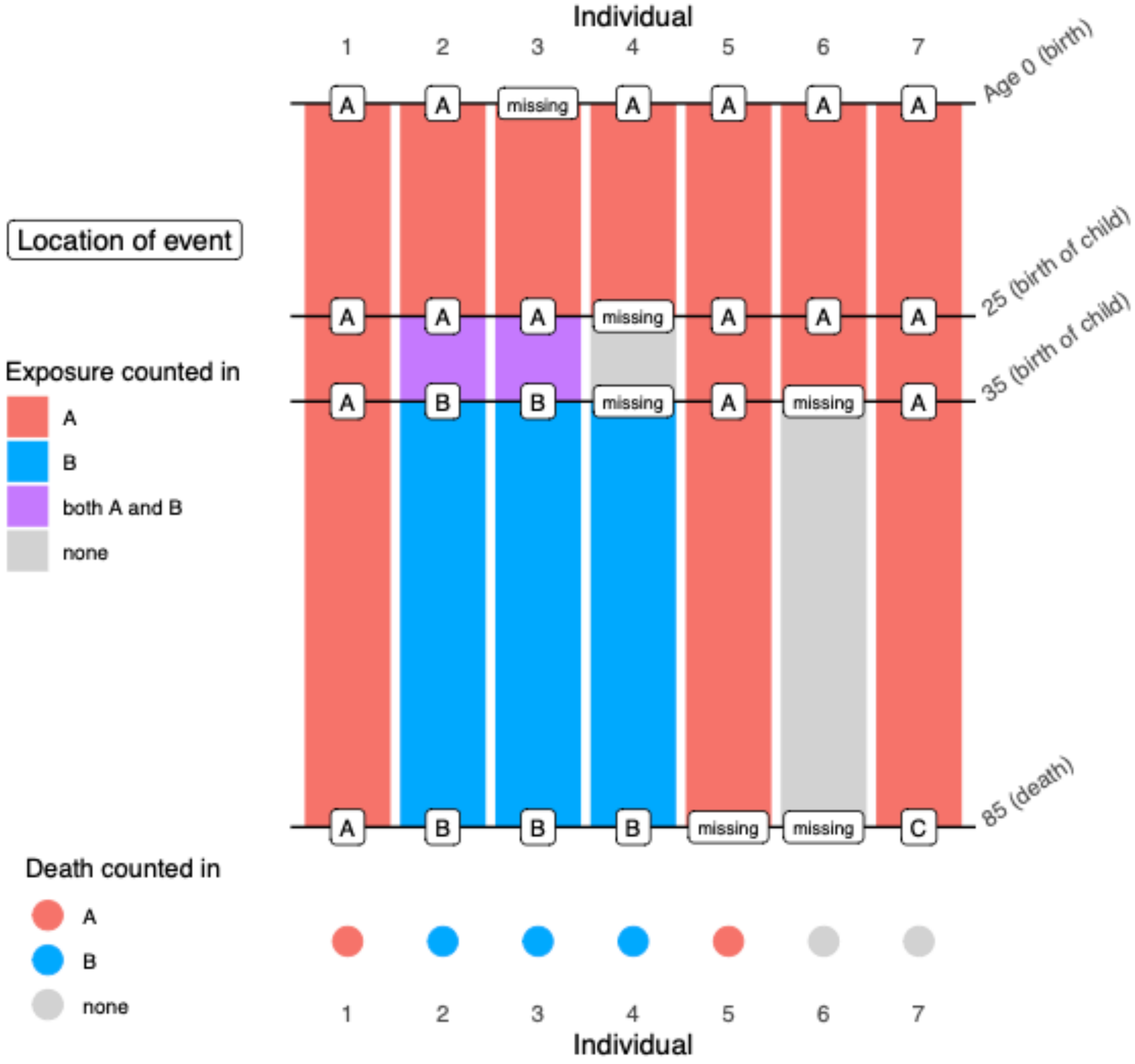
monicaalexander.com

MJAlexander

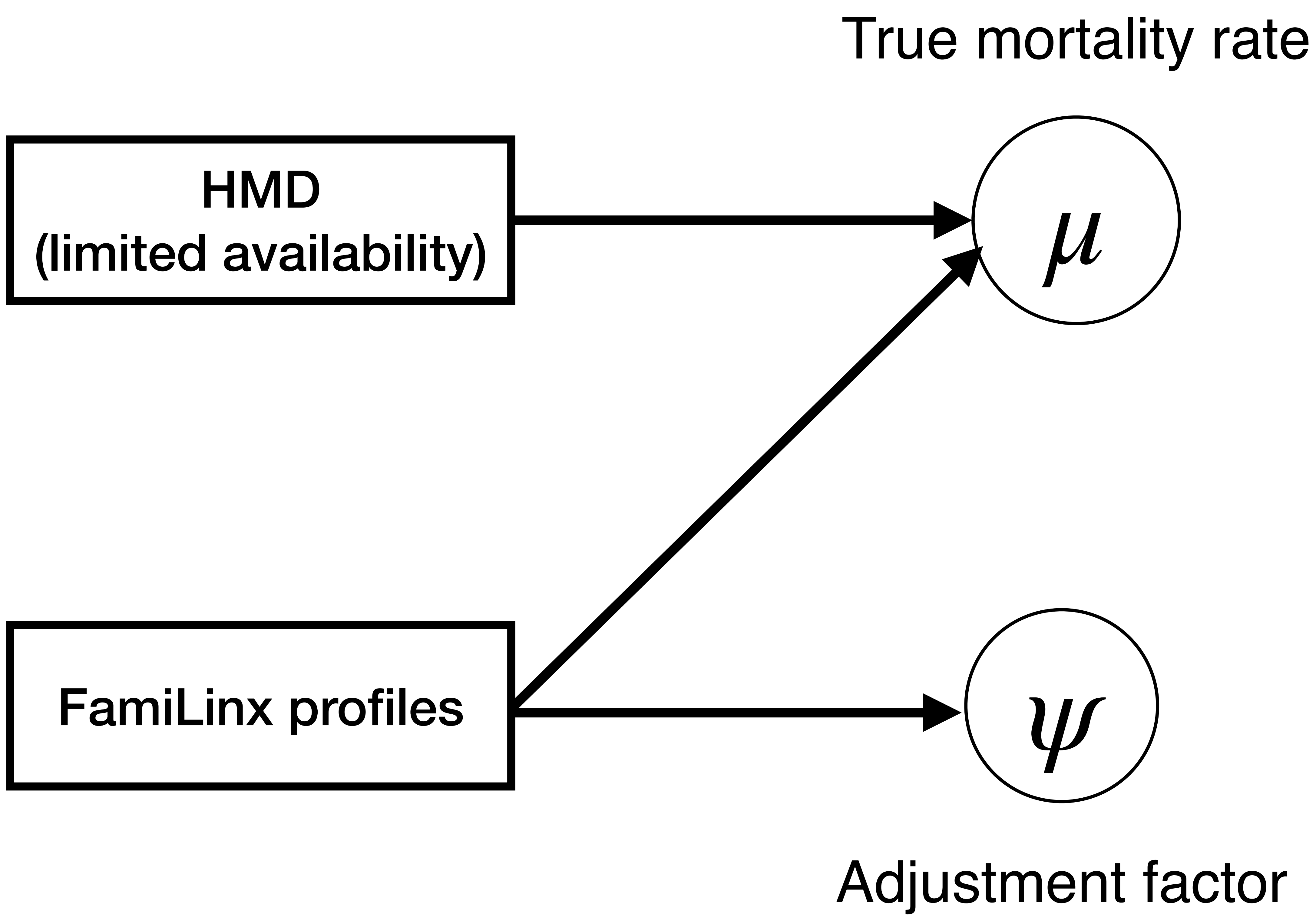
UNIVERSITY OF TORONTO

Extra

Locations


Individual
1
2
3
4
5
6
7
Age 0 (birth)
A
A
missing
A
A
A
A
Location of event
25 (birth of child)
A
A
A
missing
A
A
A
35 (birth of child)
A
B
B
missing
A
missing
A
Exposure counted in
A
B
both A and B
none
85 (death)
A
B
B
B
missing
missing
C
Death counted in
A
B
none
1
2
3
4
5
6
7
Individual

Our model compares FamiLinx-derived mortality to the Human Mortality Database (HMD)



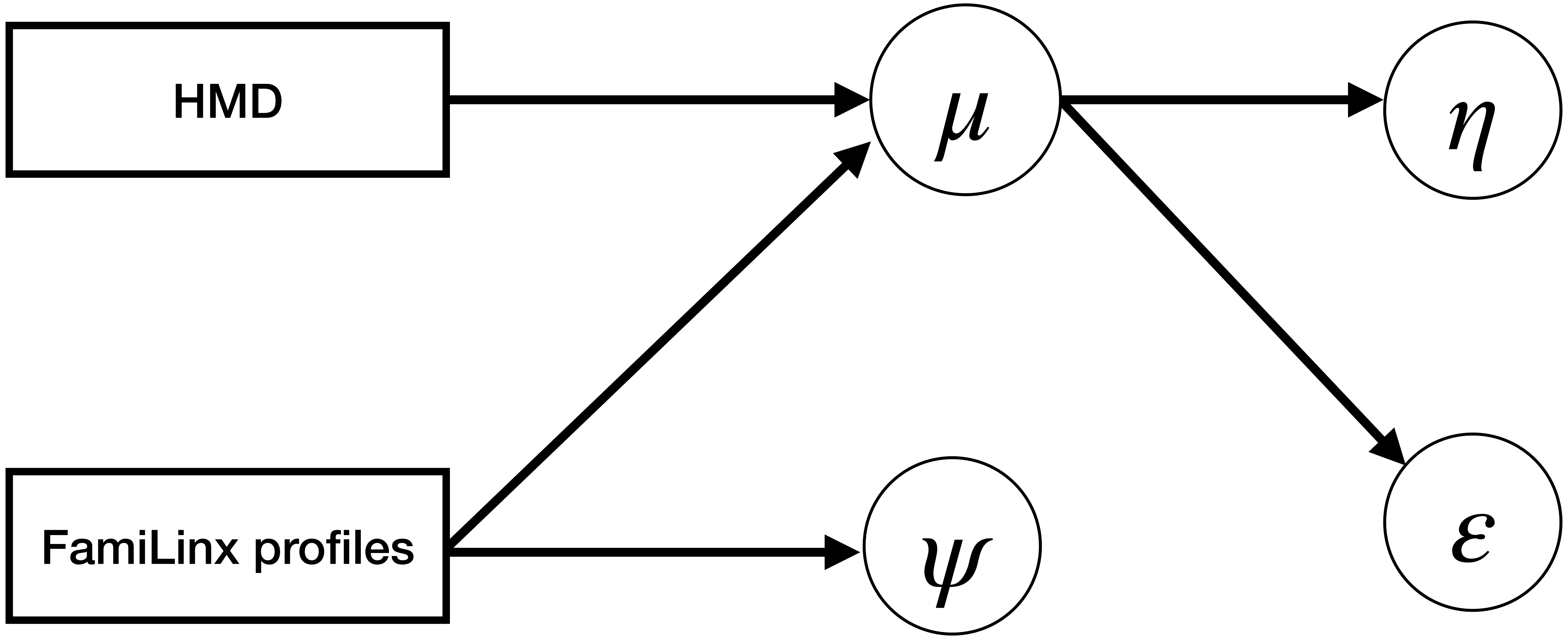
$$d^{(H)} \,|\, \mu, \phi^{(\mathrm{H})} \sim \mathrm{NegBinom}(\mu\, P^{(H)}, \phi^{(\mathrm{H})})$$

$$d^{(F)} \,|\, \mu, \psi, \phi^{(\mathrm{F})} \sim \mathrm{NegBinom}(\mu\, \psi\, P^{(F)}, \phi^{(\mathrm{F})})$$

Mortality is assumed to (mostly) follow stable patterns of age-specific mortality

True mortality rate

$$\log \vec{\mu} = \sum_{j=1}^{4} \eta_j \vec{\nu}_j + \vec{\varepsilon}$$



Adjustment factor

Systematic component

- principal components mortality model [1,2]
- time-smoothed coefficients

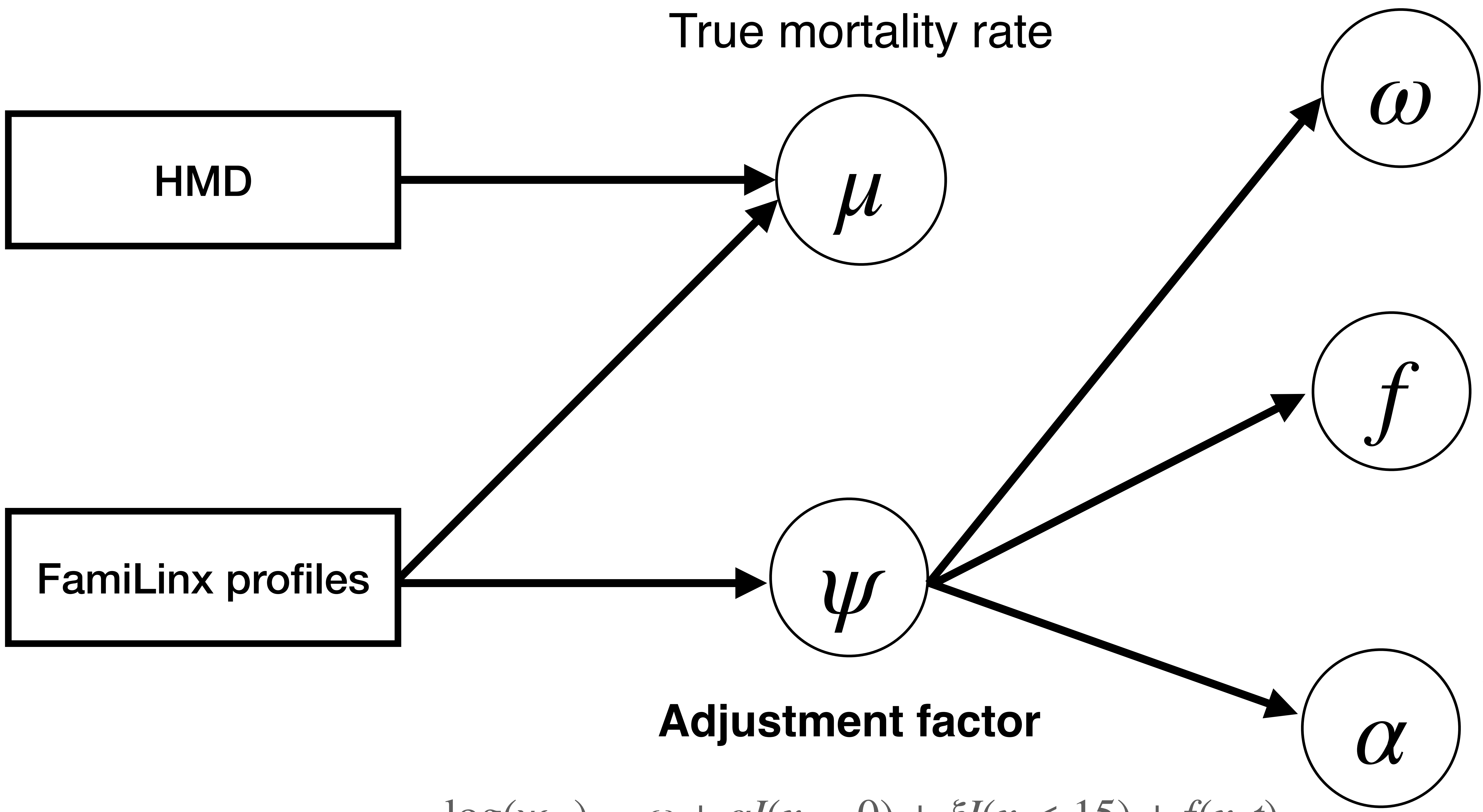
$$\eta_{j,c,g,t} = \theta_{j,c,g} + s_{j,c,g}(t)$$

Sparse deviations component

- regularized horseshoe prior [3]

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

The adjustment factors vary (mostly) smoothly over age and time



True mortality rate

HMD

FamiLinx profiles

μ

ψ

Adjustment factor

$$\log(\psi_{t,x}) = \omega + \alpha I(x = 0) + \xi I(x < 15) + f(x, t)$$

ω

Flat adjustment

- treated hierarchically (country within world)

$$\omega_c \,|\, \omega_0, \sigma_\omega \sim N(\omega_0, \sigma_\omega^2)$$

f

Smooth 2D surface over age and time

- specified as tensor product splines using a mixed-model parameterization[4]

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

α

Young age effects

- treated hierarchically (country within world)

$$\alpha_c \,|\, \alpha_0, \sigma_\alpha \sim N(\alpha_0, \sigma_\alpha^2)$$

$$\xi_c \,|\, \xi_0, \sigma_\xi \sim N(\xi_0, \sigma_\xi^2)$$

More details on tensor smooth implementation

Let $\vec{f} = \left[f\left(x_1, t_1\right), \ldots, f\left(x_N, t_N\right)\right]^T$

Reparameterize as $\vec{f}_{c,g} = X\vec{\beta}_{c,g} + Z\vec{\delta}_{c,g}$

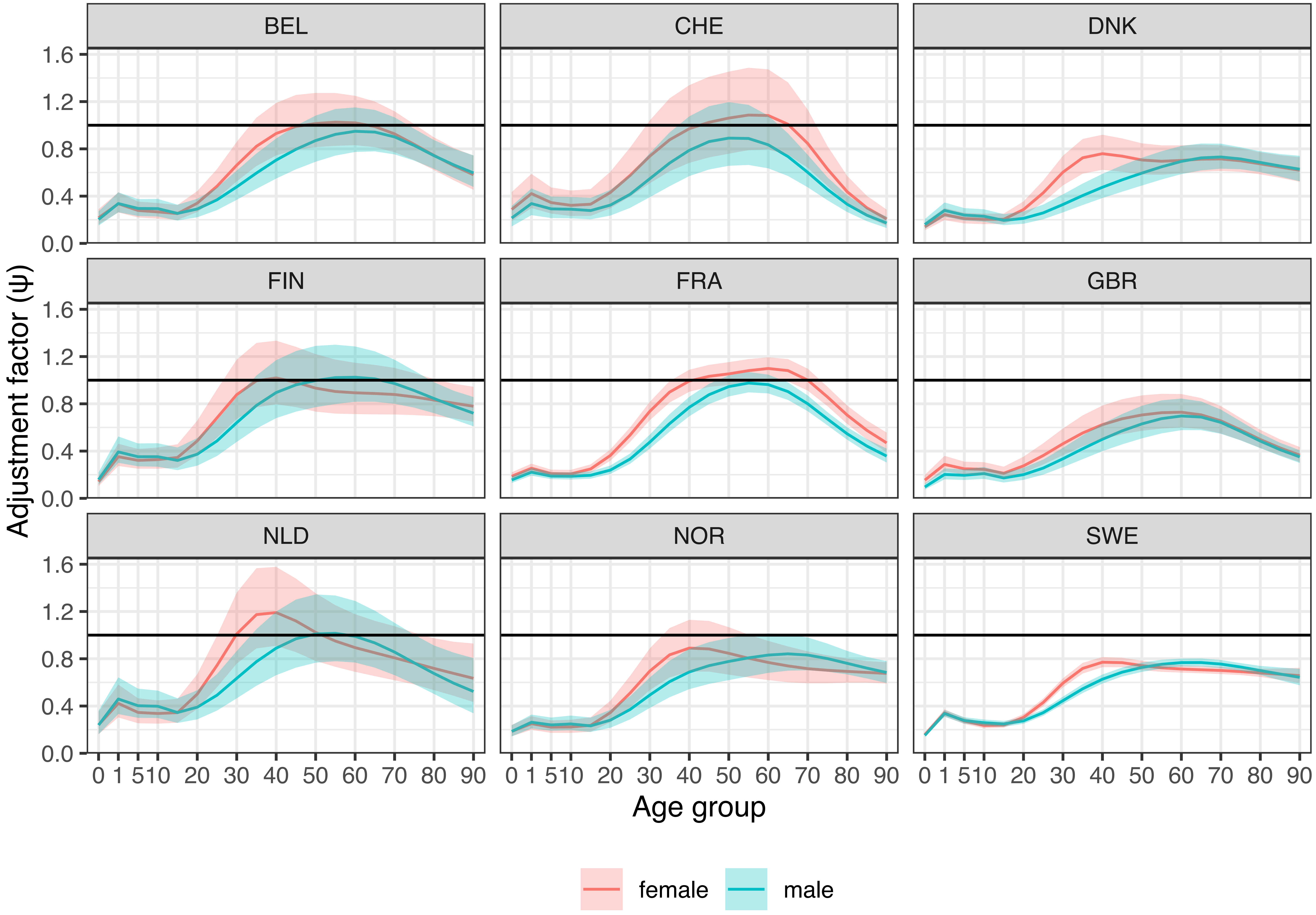
X represents unpenalized fixed effects, Z represents penalized components Using a typical 2nd derivative penalty, X contains the linear and constant functions. In the hierarchical Bayesian approach the fixed effect slopes are modelled

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

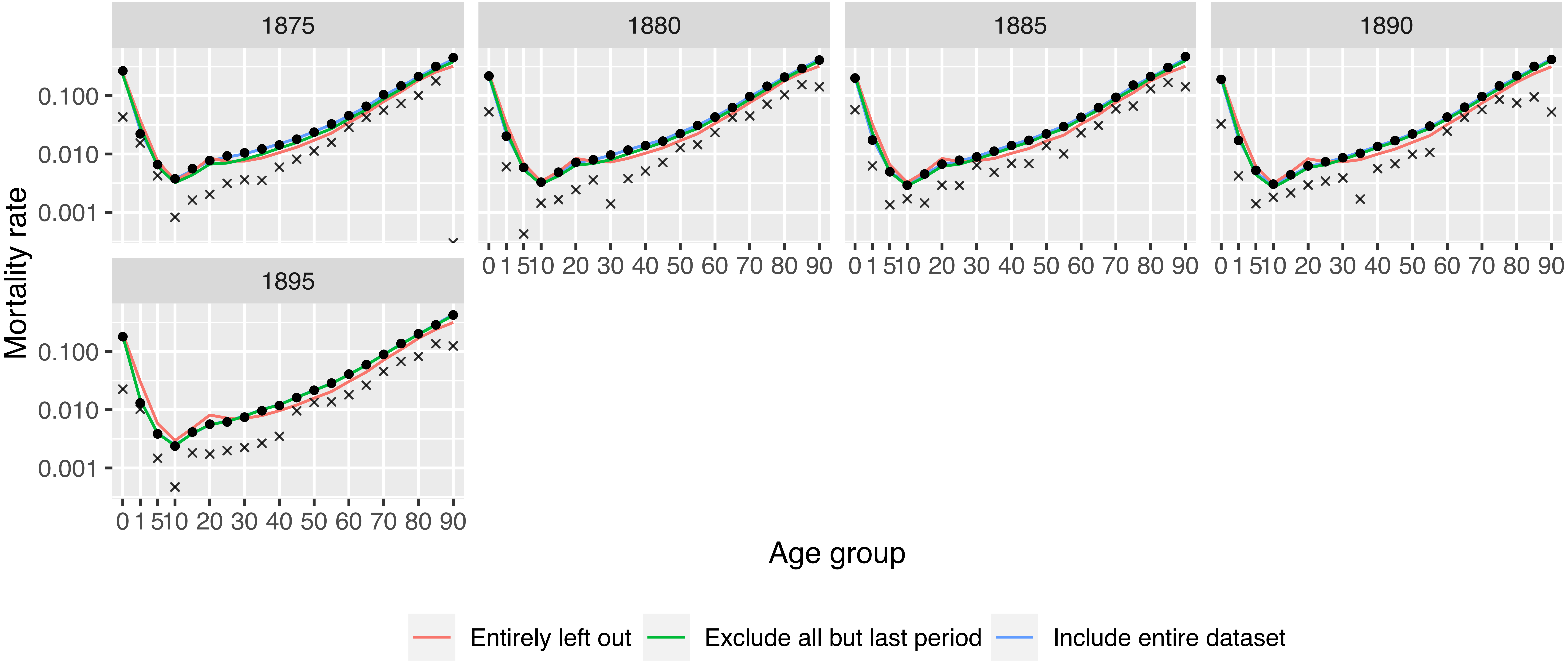
with priors $\vec{\beta}_g \sim MVN(0, I_{2\times 2})$ and $\vec{\gamma}_{c,g} \mid \vec{\sigma}_\gamma \sim MVN\left(0, \text{diag}\left(\vec{\sigma}_\gamma^2\right)\right)$.

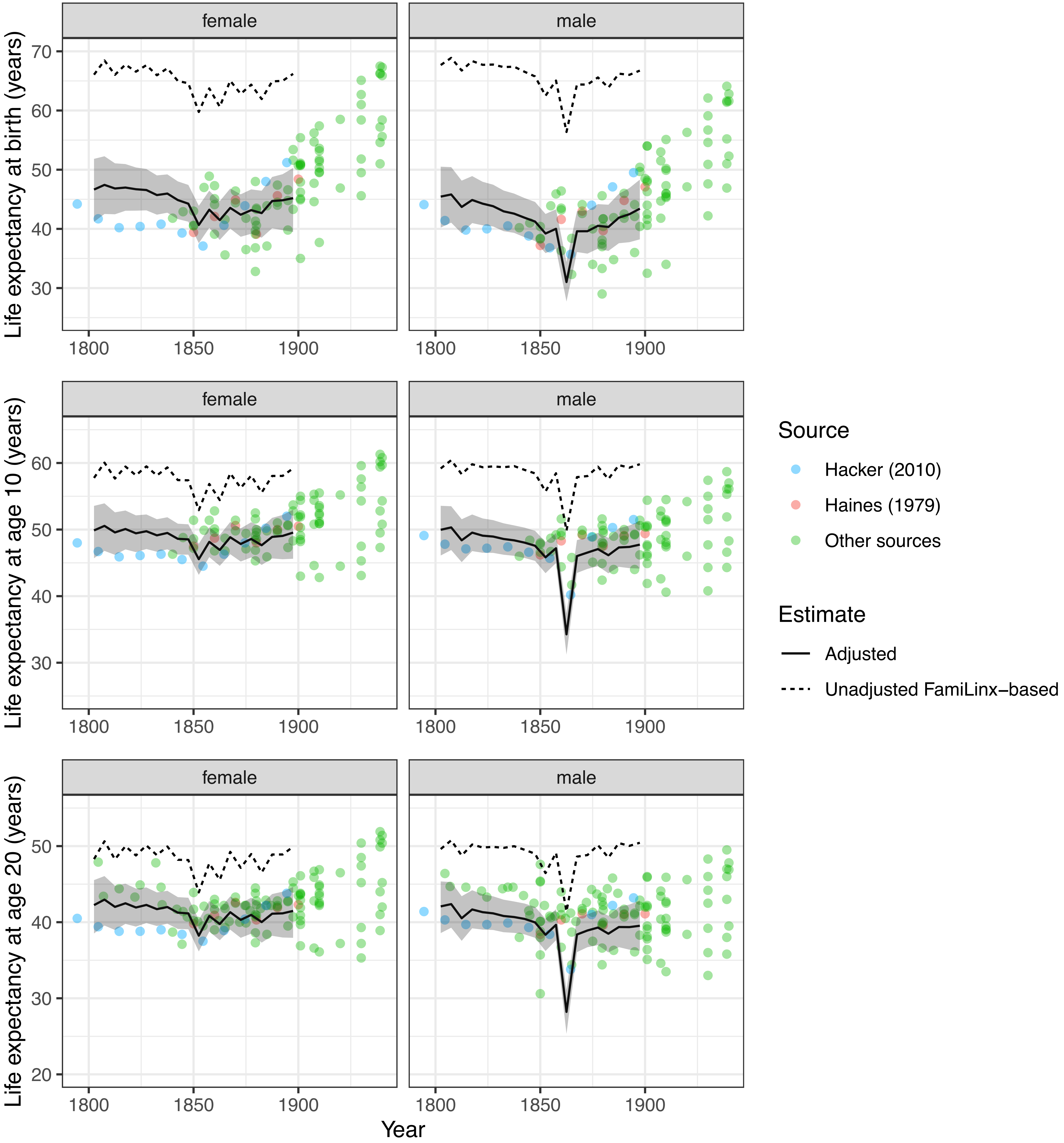
A similar approach is used for the random effects δ

Adjustment factors in 1800



Validation example




female
male
Life expectancy at birth (years)
70
60
50
40
30
1800
1850
1900
Life expectancy at age 10 (years)
60
50
40
30
Life expectancy at age 20 (years)
50
40
30
20
Year
Source
Hacker (2010)
Haines (1979)
Other sources
Estimate
Adjusted
Unadjusted FamiLinx–based