Predictive Modeling

Jonas Schöley

@jschoeley

0000-0002-3340-8518

schoeley@demogr.mpg.de

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH

Parameter-centric modeling

Tabelle 5: Cox Regression mit Interaktionseffekt (Quelle: ESA 2009, Eigene Berechnung und Darstellung)

Characteristic	HR[1]	95% CI[1]	p-value
sex			
Female	—	—	
Male	2.49	2.02, 3.07	<0.001
cohort			
_1945_1949	—	—	
_1950_1959	1.75	1.44, 2.12	<0.001
_1960_1969	1.71	1.41, 2.08	<0.001
_1970_1979	1.67	1.38, 2.03	<0.001
_1980_1991	1.71	1.42, 2.05	<0.001
bildung			
hoher_Schulab-schluss	—	—	
mittlerer_Schul-abschluss	1.39	1.29, 1.49	<0.001
niedriger_Schul-abschluss	1.49	1.36, 1.62	<0.001
Schüler_in	0.90	0.75, 1.09	0.3
sex * cohort			
Male * _1950_1959	0.60	0.46, 0.77	<0.001
Male * _1960_1969	0.57	0.44, 0.74	<0.001
Male * _1970_1979	0.52	0.40, 0.67	<0.001
Male * _1980_1991	0.47	0.37, 0.59	<0.001

[1]HR = Hazard Ratio, CI = Confidence Interval

Balling (2022). Das Rauchverhalten von Männern und Frauen in verschiedenen Kohorten.

Parameter-centric modeling vs. predictive modeling

Balling (2022). Das Rauchverhalten von Männern und Frauen in verschiedenen Kohorten.

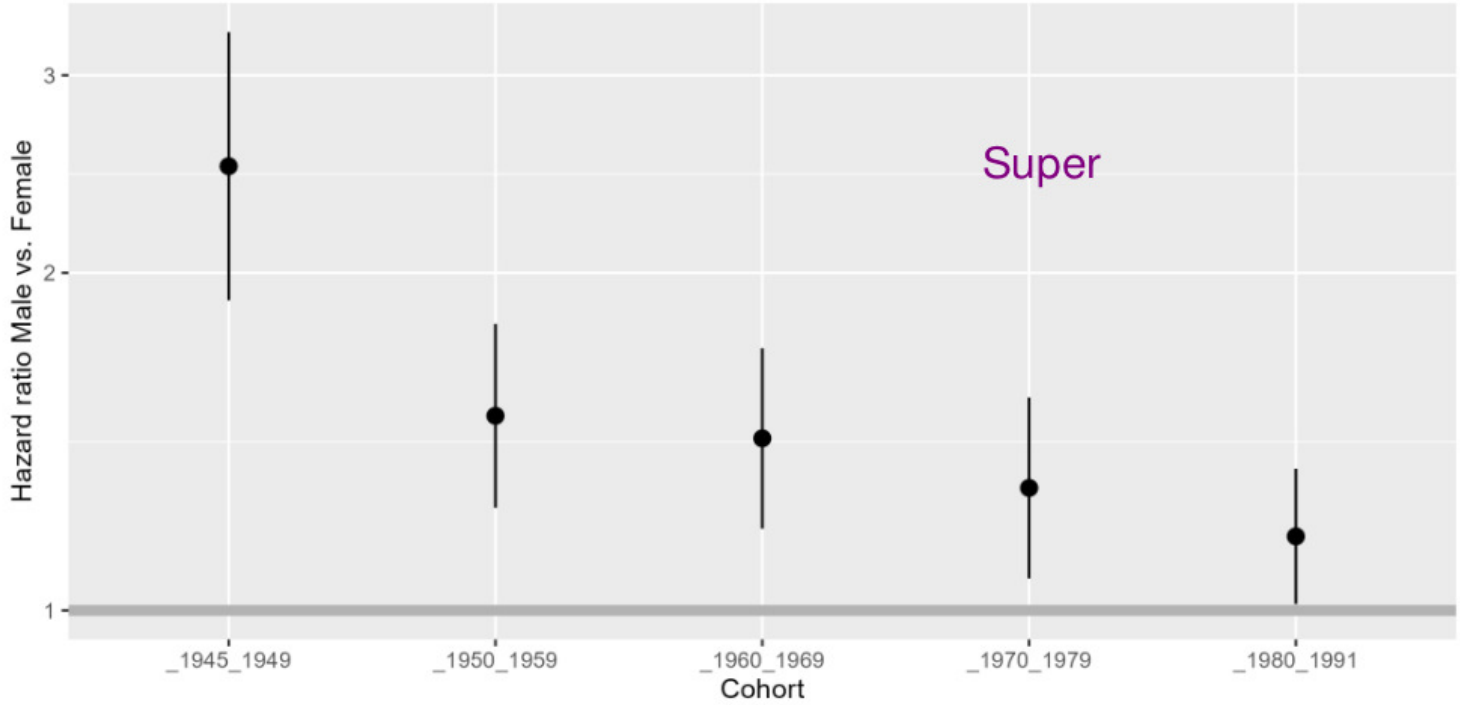
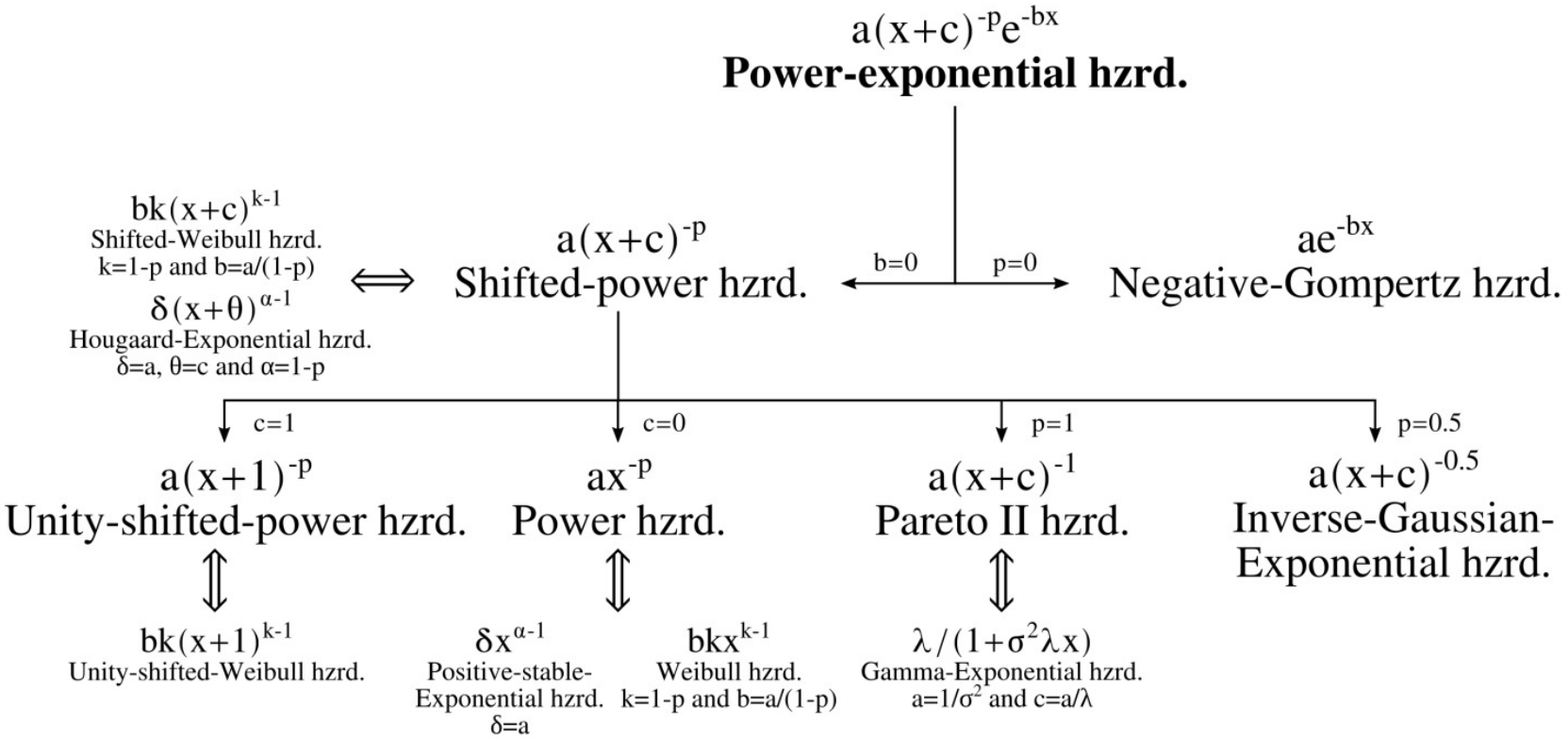


ABBILDUNG 6: INTERAKTIONSEFFEKT (QUELLE: ESA 2009, EIGENE BERECHNUNG UND DARSTELLUNG)

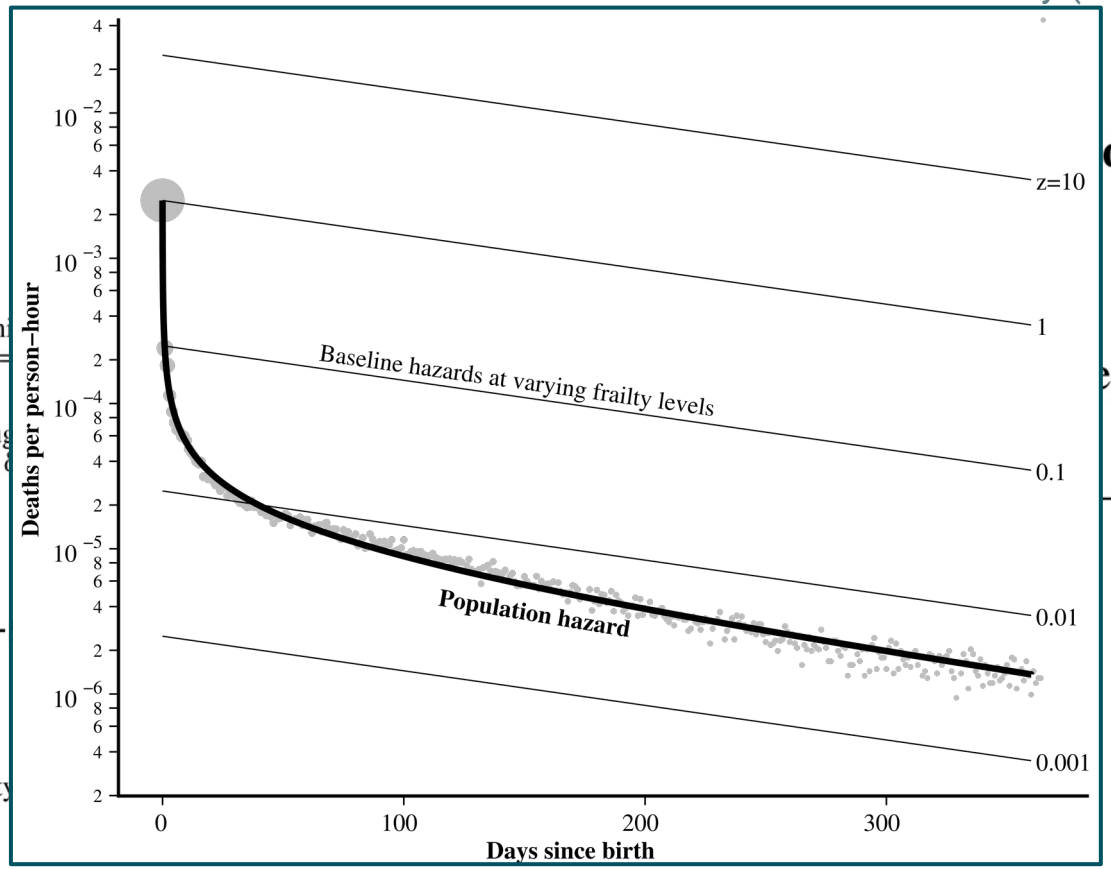
Parameter-centric modeling vs. predictive modeling

Schöley (2020). The dynamics of ontogenescence. Github link.



Parameter-centric modeling vs. predictive modeling

Schöley (2020). The dynamics of ontogenescence. Github link.



ae^{-bx}

egative-Gompertz hzrd.

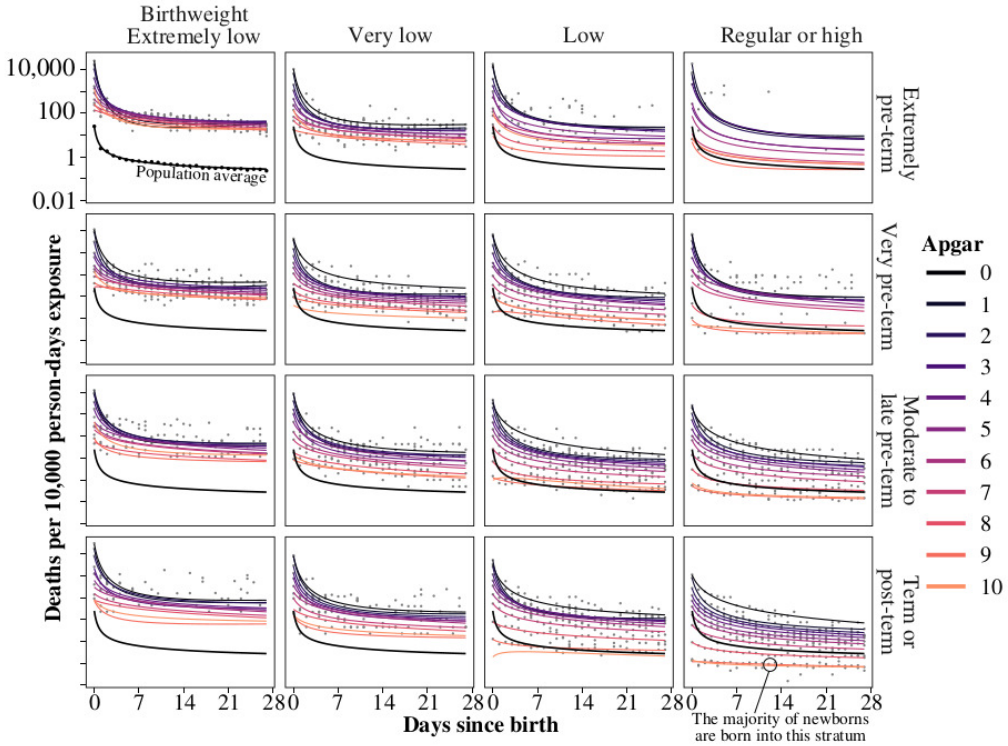
p=0.5

$a(x+c)^{-0.5}$

Inverse-Gaussian-Exponential hzrd.

Parameter-centric modeling vs. predictive modeling

Figure 3: Estimated hazard rates versus life table mortality rates over age by prematurity, birth weight, and Apgar score. There is substantial heterogeneity in the level and the shape of the neonatal hazard trajectories ruling out the hypothesis of proportional frailties. The black line in each panel shows the estimated hazard trajectory for the entire birth cohort.



Schöley (2020). The dynamics of ontogenescence. Github link.

Parameter-centric modeling vs. predictive modeling

Figure 3: Estimated hazard rates versus life table mortality rates over age by prematurity, birth weight, and Apgar score. There is substantial heterogeneity in the level and the shape of the neonatal hazard trajectories ruling out the hypothesis of proportional frailties. The black line in each panel shows the estimated hazard trajectory for the entire birth cohort.

Schöley (2020). The dynamics of ontogenescence. Github link.

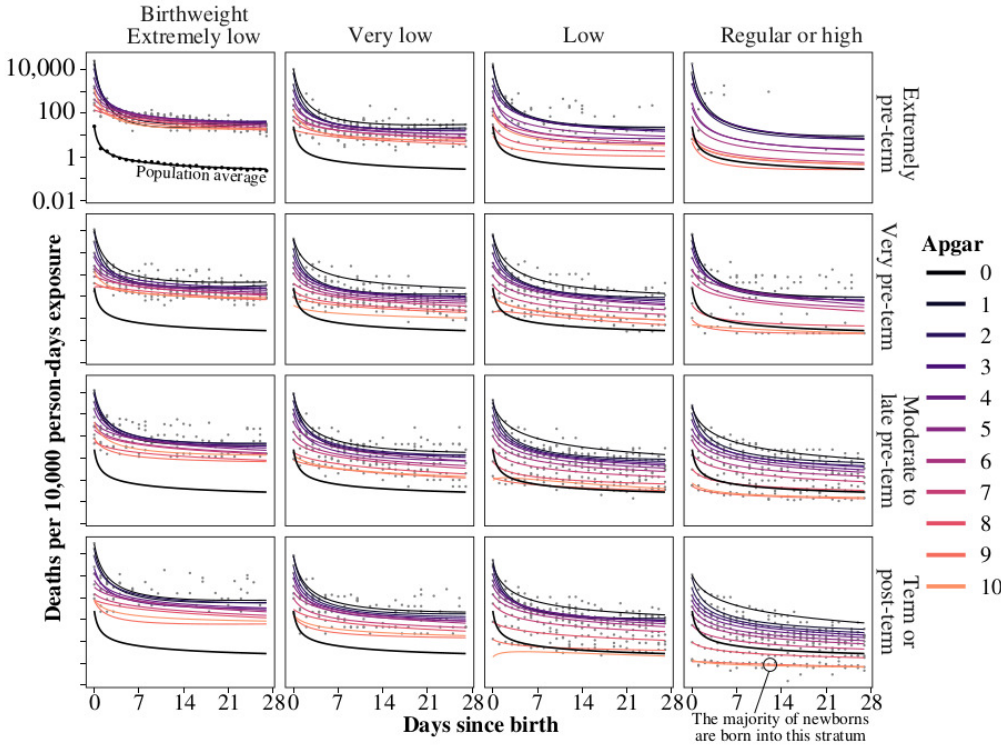
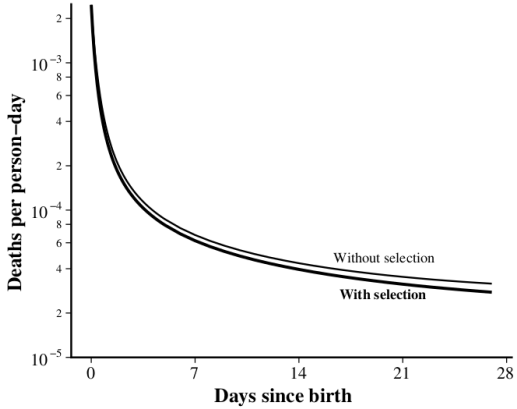
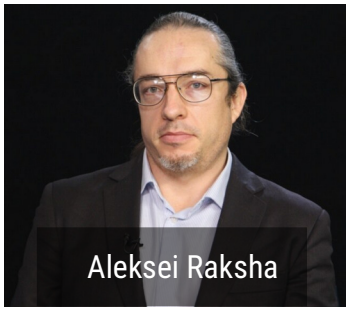


Figure 4: Keeping the population composition fixed at the distribution observed at birth only results in a minor change of the population hazard trajectory.



Predictive modeling in demography: Indirect estimation



Rogue Demographer Says Russia May Top Europe in Covid Deaths

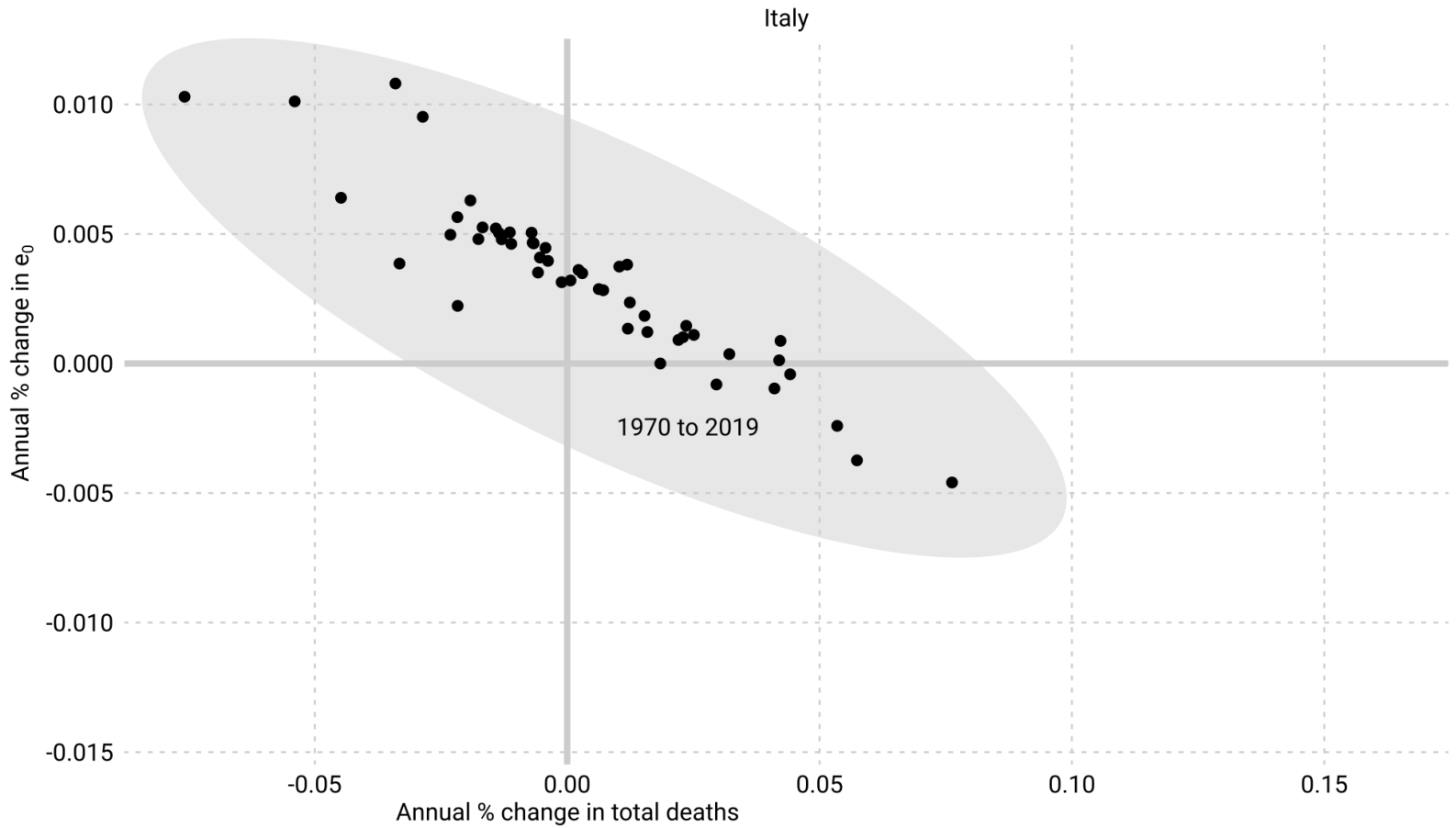
- Former statistics agency employee has criticized virus data
- Daily death numbers need to be tripled, demographer says

Photographer: Dimitar Dilkoff/AFP via Getty Images

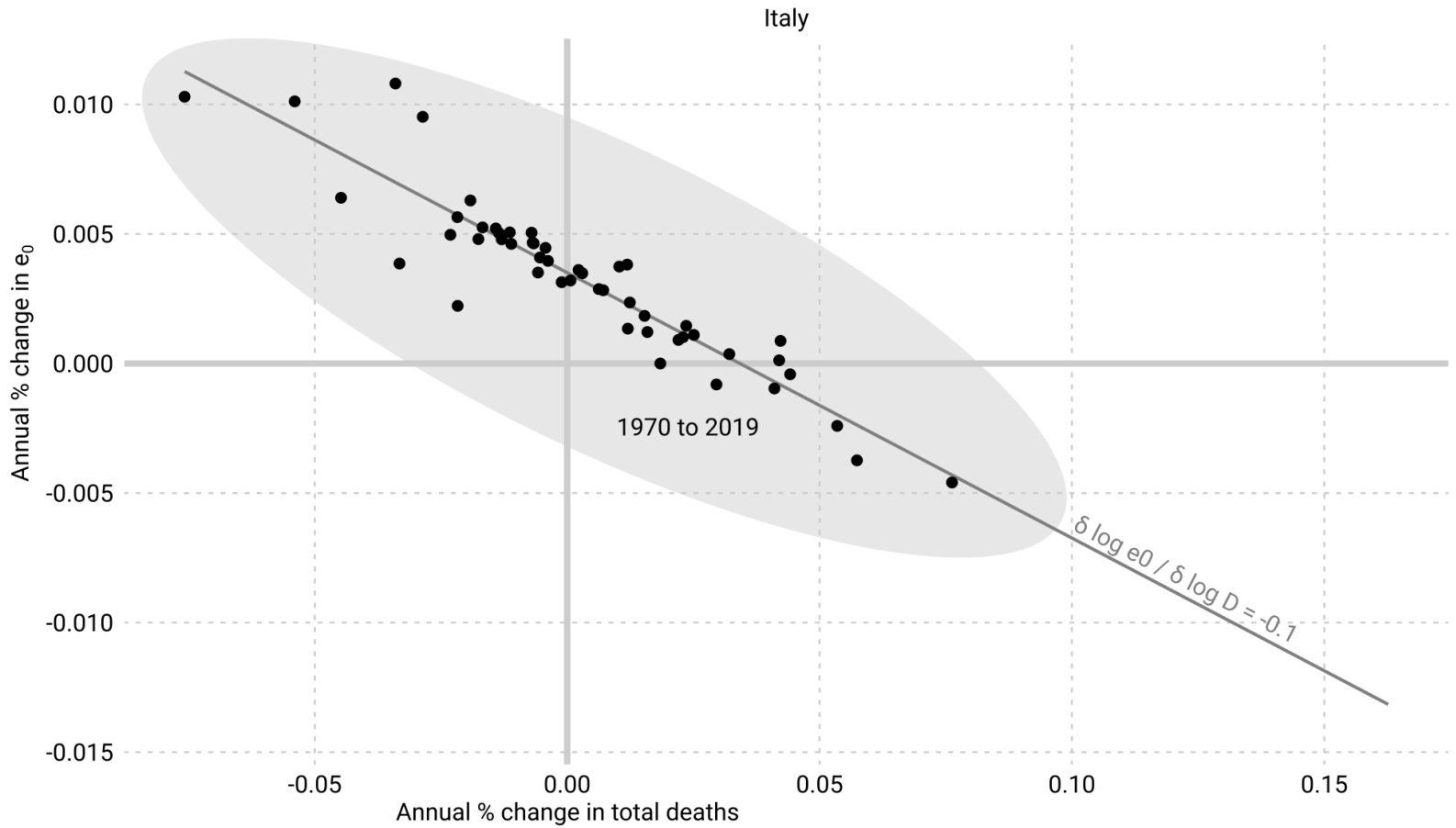
By Anya Andrianova and Stepan Kravchenko
October 20, 2020 at 6:00 AM GMT+2 *Updated on October 20, 2020 at 12:36 PM GMT+2*

Bloomberg. October 20, 2020. Link.

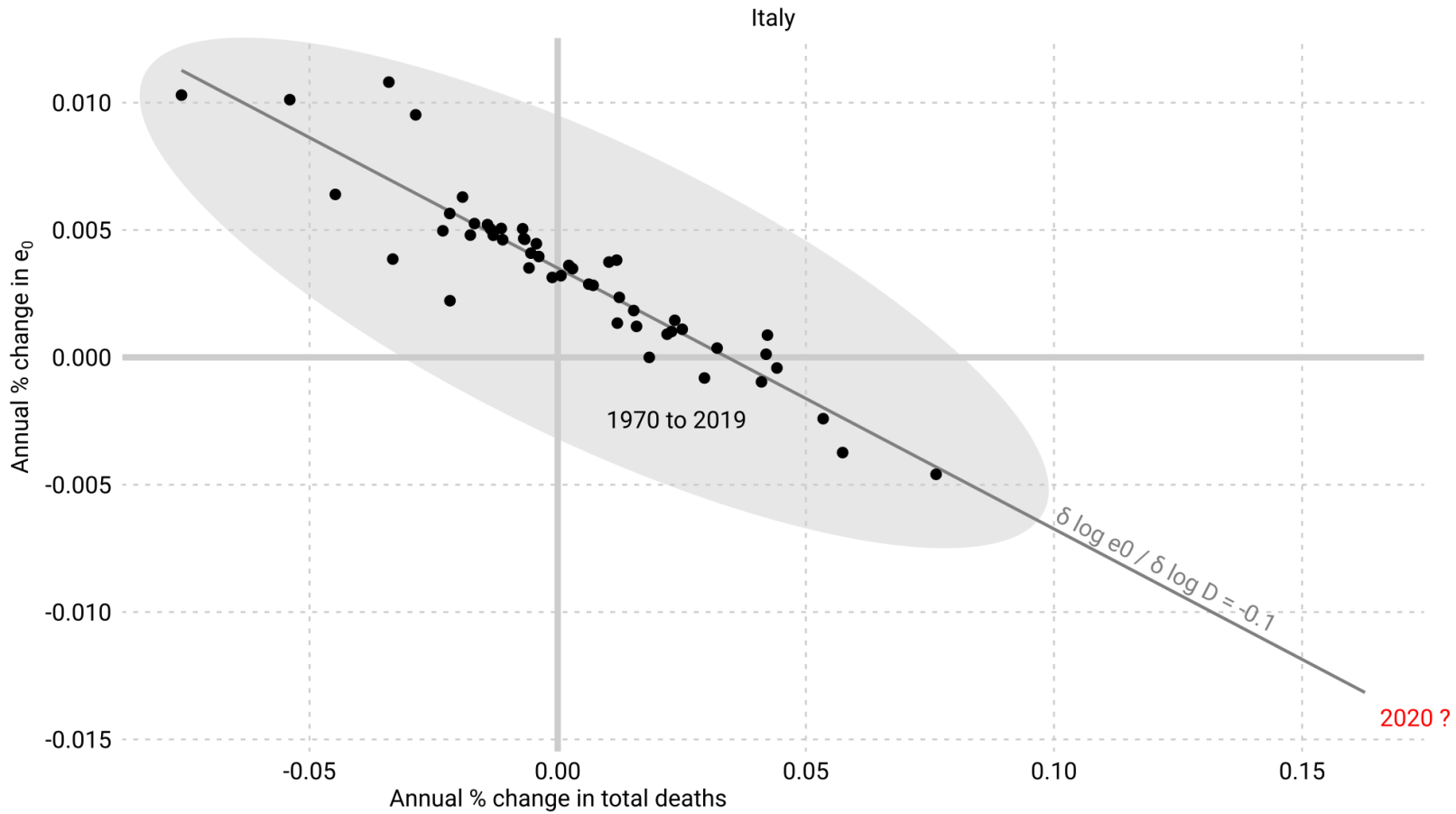
Predictive modeling in demography: Indirect estimation



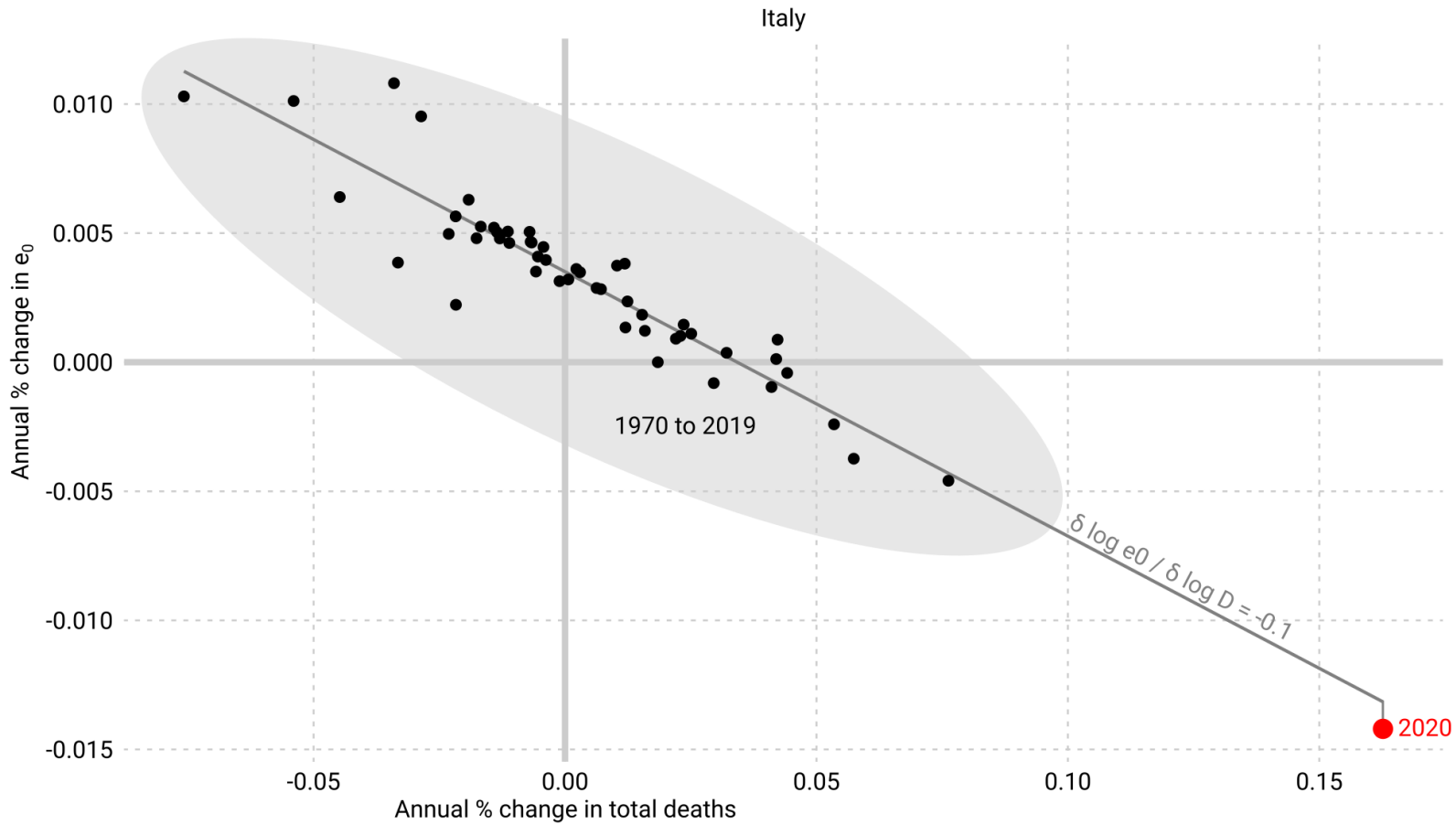
Predictive modeling in demography: Indirect estimation



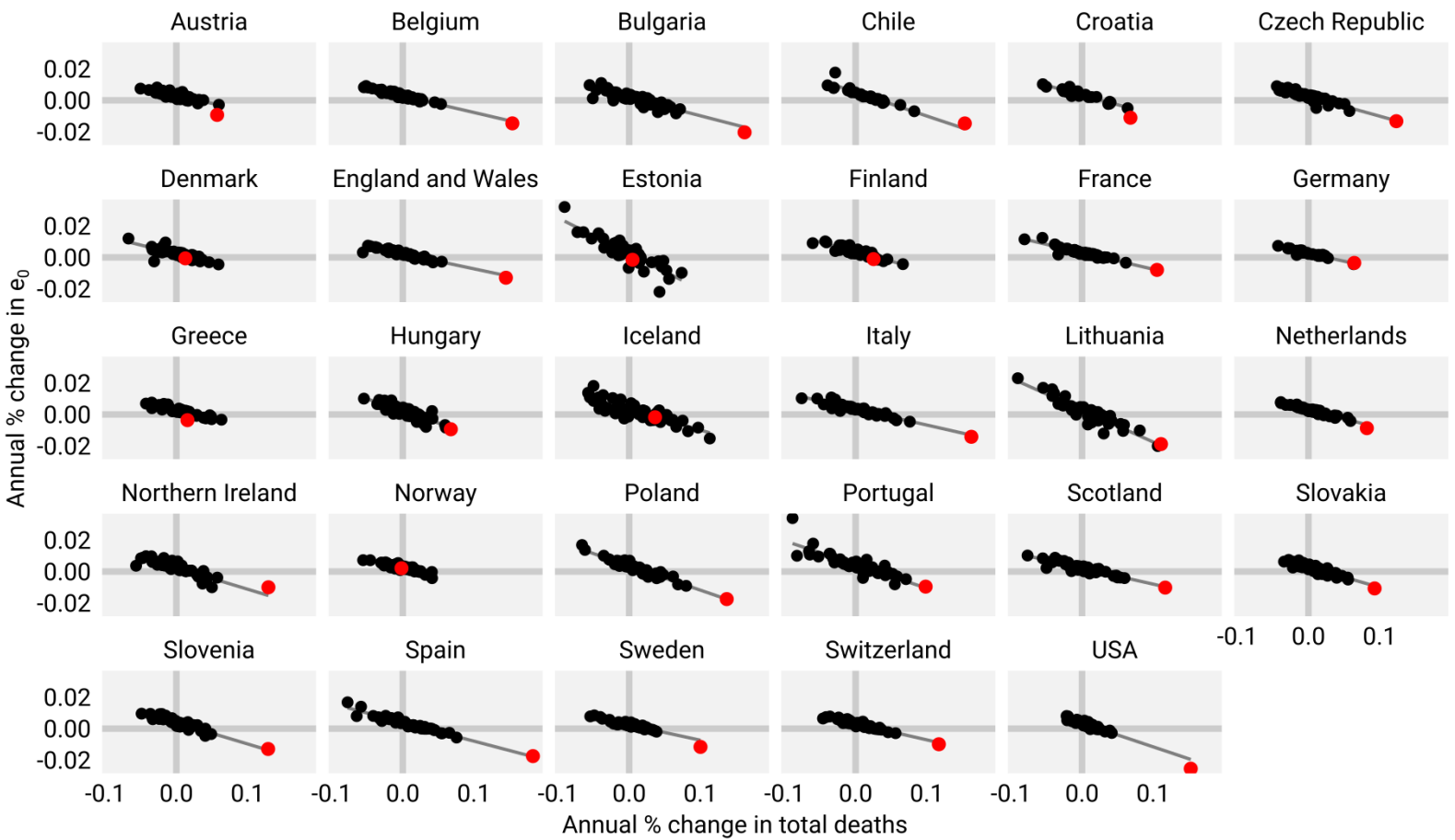
Predictive modeling in demography: Indirect estimation



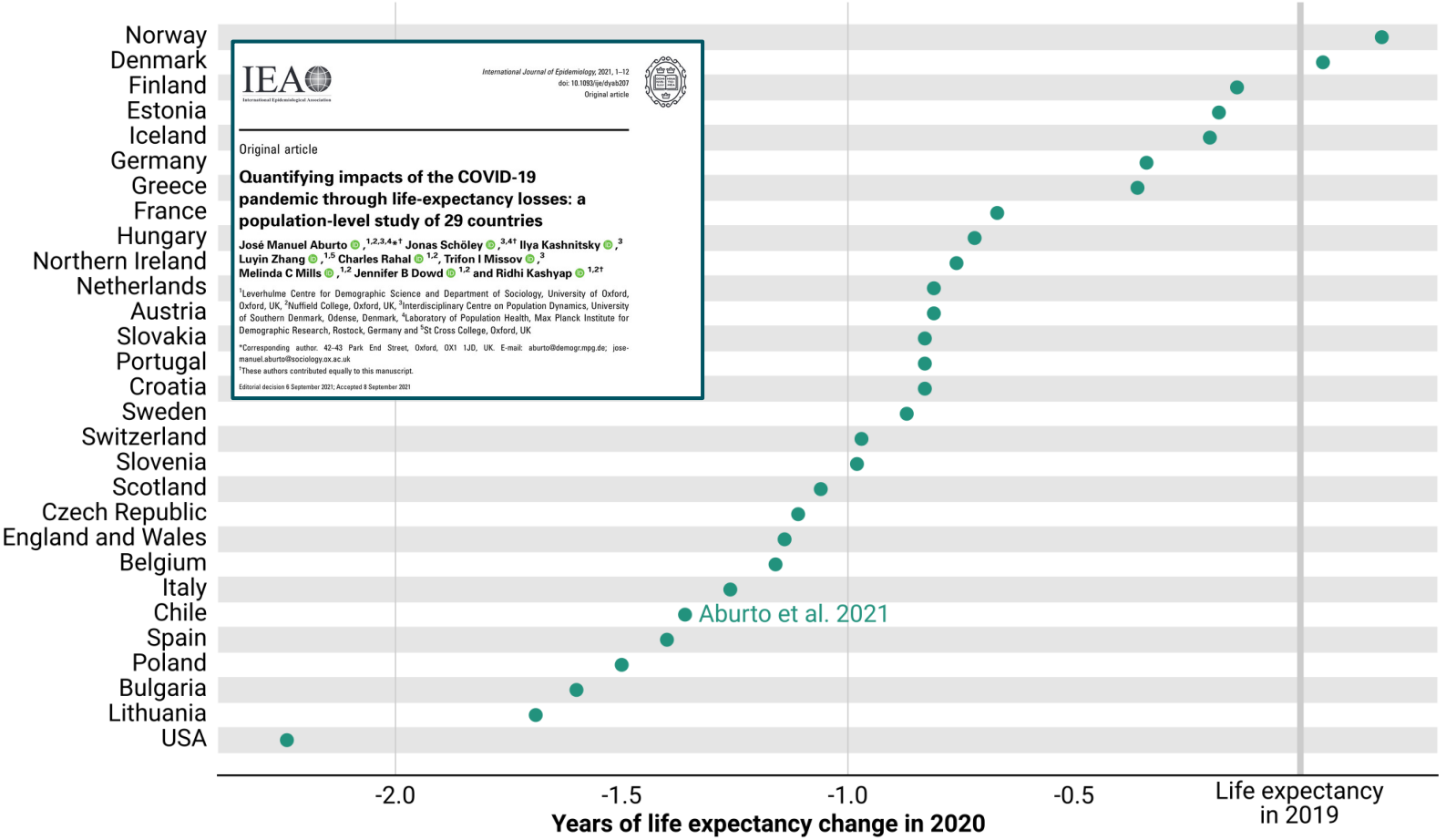
Predictive modeling in demography: Indirect estimation



Predictive modeling in demography: Indirect estimation



Predictive modeling in demography: Indirect estimation



IEA International Epidemiological Association

International Journal of Epidemiology, 2021, 1–12
doi: 10.1093/ije/dyab207
Original article

Original article

Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries

José Manuel Aburto [1,2,3,4*†] Jonas Schöley [3,4†] Ilya Kashnitsky [3] Luyin Zhang [1,5] Charles Rahal [1,2], Trifon I Missov [3] Melinda C Mills [1,2] Jennifer B Dowd [1,2] and Ridhi Kashyap [1,2†]

[1]Leverhulme Centre for Demographic Science and Department of Sociology, University of Oxford, Oxford, UK, [2]Nuffield College, Oxford, UK, [3]Interdisciplinary Centre on Population Dynamics, University of Southern Denmark, Odense, Denmark, [4]Laboratory of Population Health, Max Planck Institute for Demographic Research, Rostock, Germany and [5]St Cross College, Oxford, UK

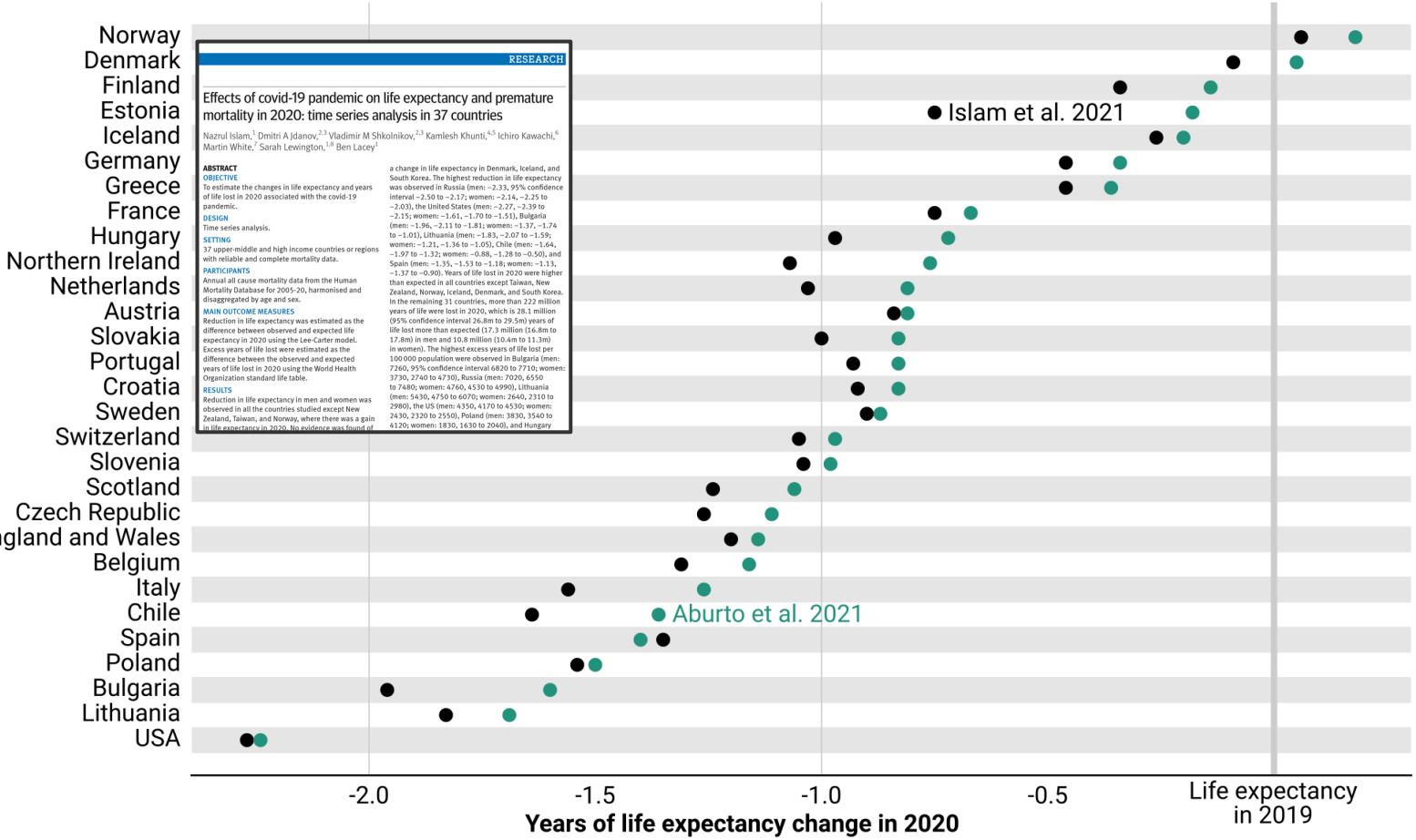
*Corresponding author. 42–43 Park End Street, Oxford, OX1 1JD, UK. E-mail: aburto@demogr.mpg.de; jose-manuel.aburto@sociology.ox.ac.uk

†These authors contributed equally to this manuscript.

Editorial decision 6 September 2021; Accepted 8 September 2021

Aburto & Schöley et al. 2021. Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries. 10.1093/ije/dyab207

Predictive modeling in demography: Indirect estimation



RESEARCH

Effects of covid-19 pandemic on life expectancy and premature mortality in 2020: time series analysis in 37 countries

Nazrul Islam,[1] Dmitri A Jdanov,[2,3] Vladimir M Shkolnikov,[2,3] Kamlesh Khunti,[4,5] Ichiro Kawachi,[6] Martin White,[7] Sarah Lewington,[1,8] Ben Lacey[1]

ABSTRACT

OBJECTIVE
To estimate the changes in life expectancy and years of life lost in 2020 associated with the covid-19 pandemic.

DESIGN
Time series analysis.

SETTING
37 upper-middle and high income countries or regions with reliable and complete mortality data.

PARTICIPANTS
Annual all cause mortality data from the Human Mortality Database for 2005-20, harmonised and disaggregated by age and sex.

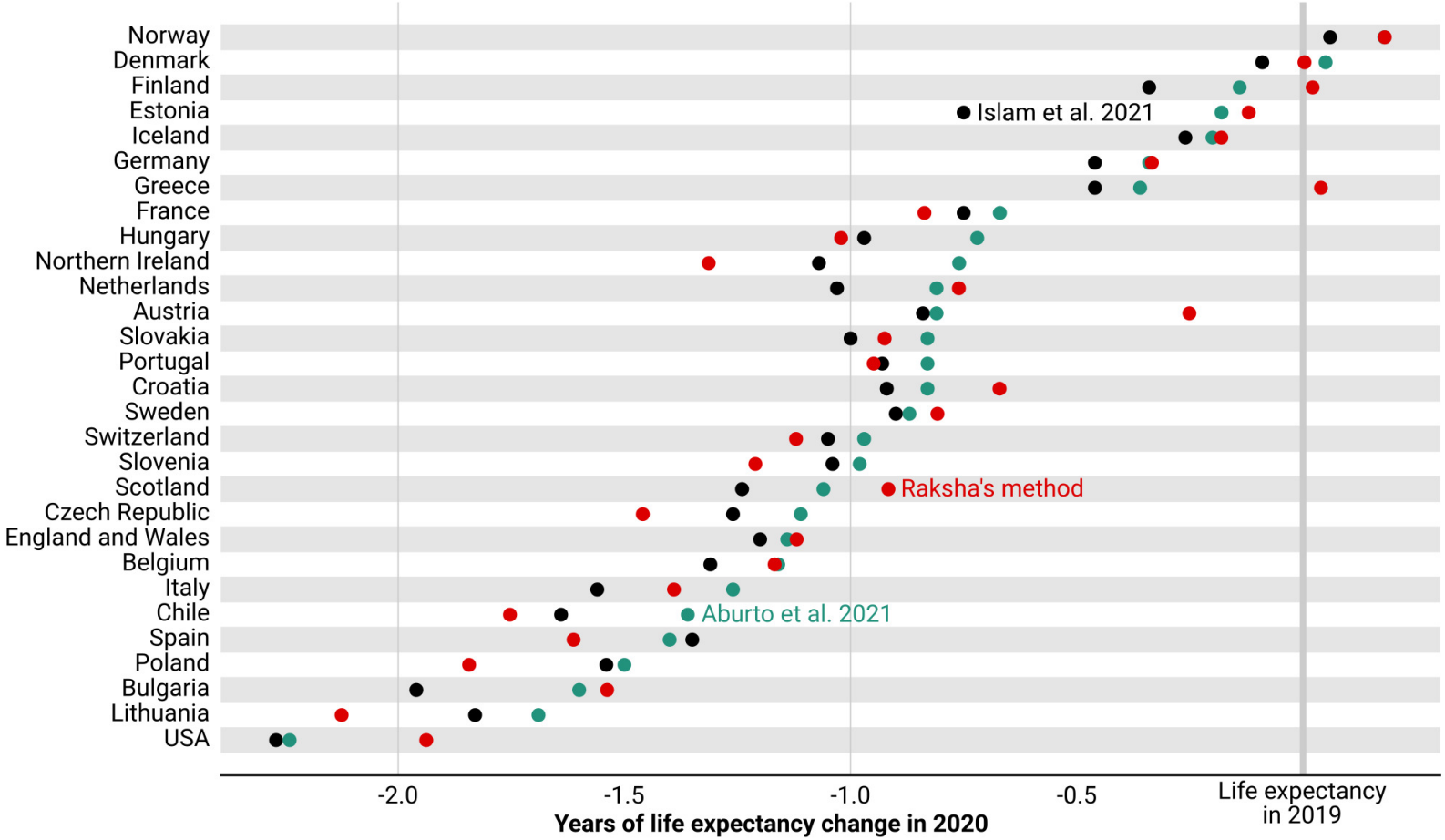
MAIN OUTCOME MEASURES
Reduction in life expectancy was estimated as the difference between observed and expected life expectancy in 2020 using the Lee-Carter model. Excess years of life lost were estimated as the difference between the observed and expected years of life lost in 2020 using the World Health Organization standard life table.

RESULTS
Reduction in life expectancy in men and women was observed in all the countries studied except New Zealand, Taiwan, and Norway, where there was a gain in life expectancy in 2020. No evidence was found of a change in life expectancy in Denmark, Iceland, and South Korea. The highest reduction in life expectancy was observed in Russia (men: −2.33, 95% confidence interval −2.50 to −2.17; women: −2.14, −2.25 to −2.03), the United States (men: −2.27, −2.39 to −2.15; women: −1.61, −1.70 to −1.51), Bulgaria (men: −1.96, −2.11 to −1.81; women: −1.37, −1.74 to −1.01), Lithuania (men: −1.83, −2.07 to −1.59; women: −1.21, −1.36 to −1.05), Chile (men: −1.64, −1.97 to −1.32; women: −0.88, −1.28 to −0.50), and Spain (men: −1.35, −1.53 to −1.18; women: −1.13, −1.37 to −0.90). Years of life lost in 2020 were higher than expected in all countries except Taiwan, New Zealand, Norway, Iceland, Denmark, and South Korea. In the remaining 31 countries, more than 222 million years of life were lost in 2020, which is 28.1 million (95% confidence interval 26.8m to 29.5m) years of life lost more than expected (17.3 million (16.8m to 17.8m) in men and 10.8 million (10.4m to 11.3m) in women). The highest excess years of life lost per 100 000 population were observed in Bulgaria (men: 7260, 95% confidence interval 6820 to 7710; women: 3730, 2740 to 4730), Russia (men: 7020, 6550 to 7480; women: 4760, 4530 to 4990), Lithuania (men: 5430, 4750 to 6070; women: 2640, 2310 to 2980), the US (men: 4350, 4170 to 4530; women: 2430, 2320 to 2550), Poland (men: 3830, 3540 to 4120; women: 1830, 1630 to 2040), and Hungary

Aburto & Schöley et al. 2021. Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries. 10.1093/ije/dyab207

Islam & Jdanov et al. 2021. Effects of covid-19 pandemic on life expectancy and premature mortality in 2020: time series analysis in 37 countries. 10.1136/bmj-2021-066768

Predictive modeling in demography: Indirect estimation



Aburto & Schöley et al. 2021. Quantifying impacts of the COVID-19 pandemic through life-expectancy losses: a population-level study of 29 countries. 10.1093/ije/dyab207

Islam & Jdanov et al. 2021. Effects of covid-19 pandemic on life expectancy and premature mortality in 2020: time series analysis in 37 countries. 10.1136/bmj-2021-066768

Predictive modeling in demography: Indirect estimation

Kashnitsky, Raksha, Schöley etal. (2021). Monitoring life expectancy. Link to video

Predictive modeling in demography: Spatial analyses

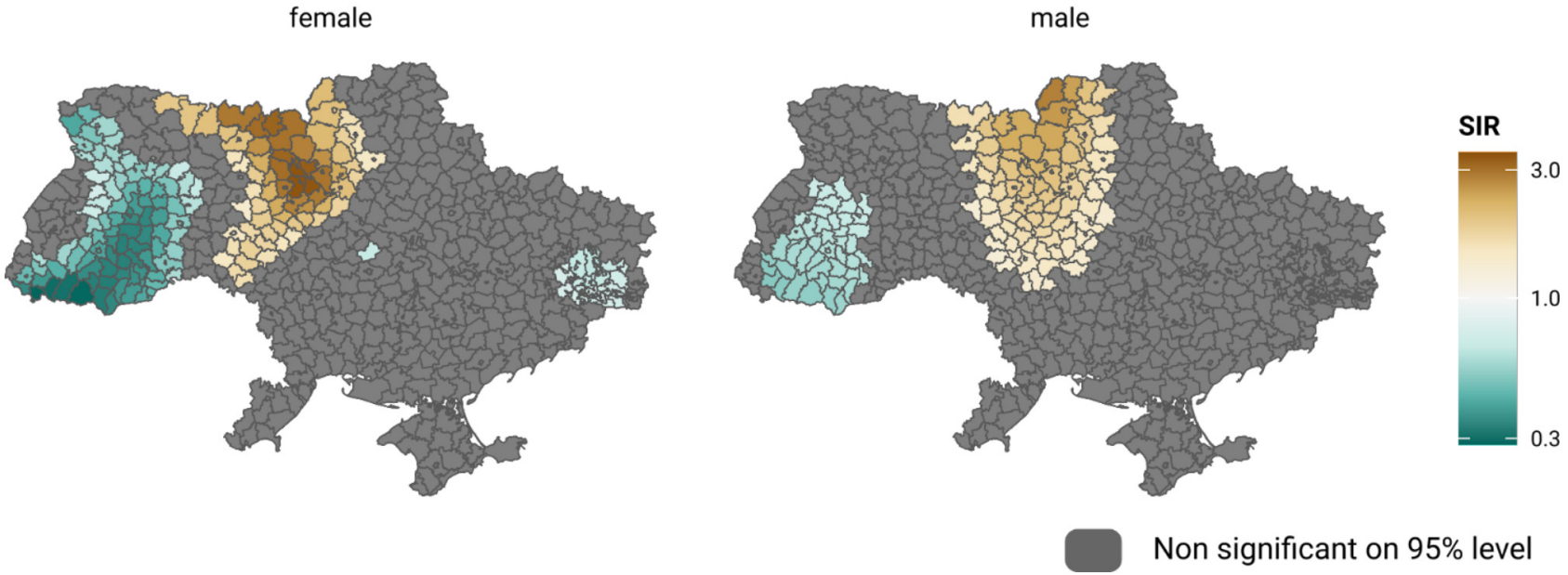
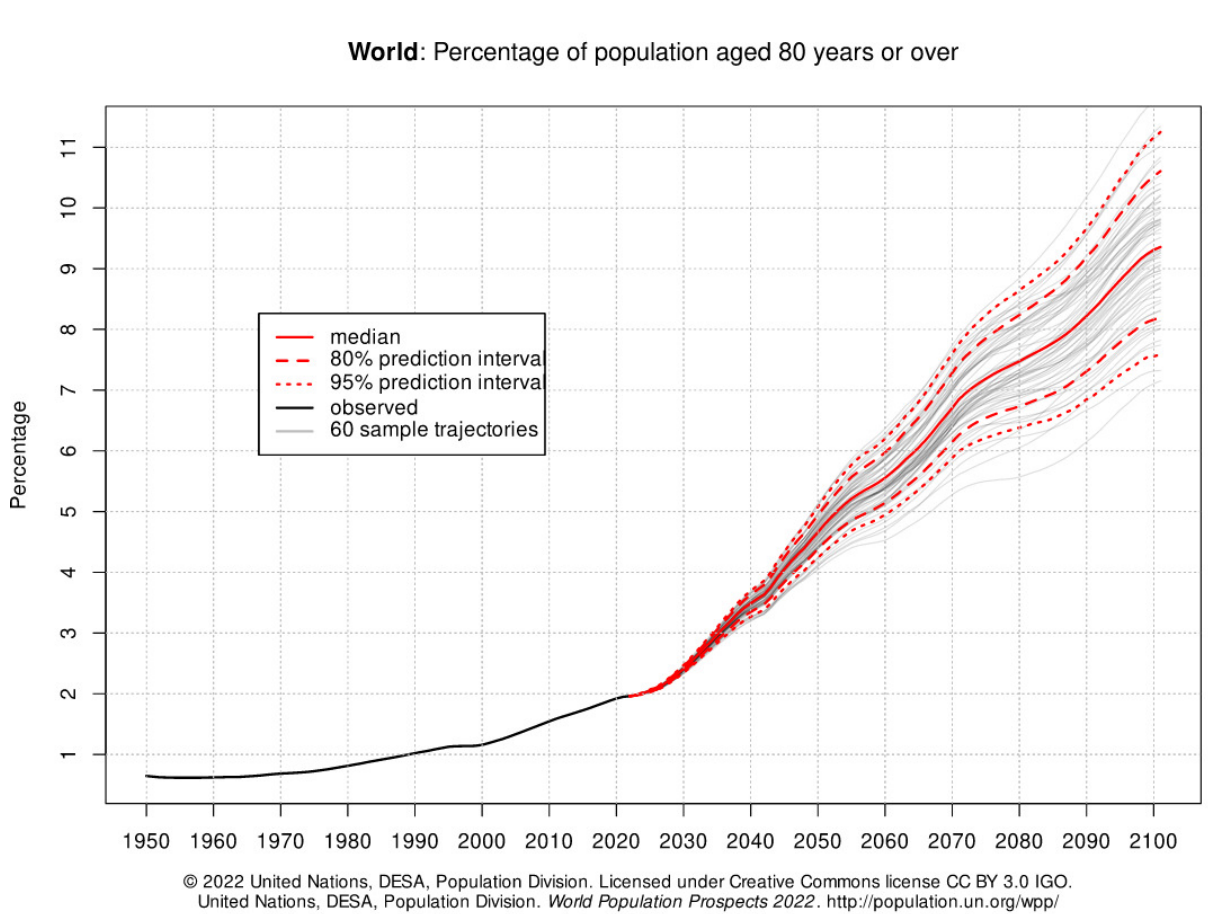


Figure 1 Standardized incidence ratio of thyroid cancer in Ukraine 2001 for the population aged 15+. The standard is given by the average district-level incidence. These are estimates based on an Overdispersed Poisson Regression with 2D P-splines for a smooth nonparametric estimate of the spatial effect.

Levchuk, Cilek, Schöley, Jasilionis (2022). Long-term health effects of the Chernobyl accident.

Predictive modeling in demography: Projections and forecasts

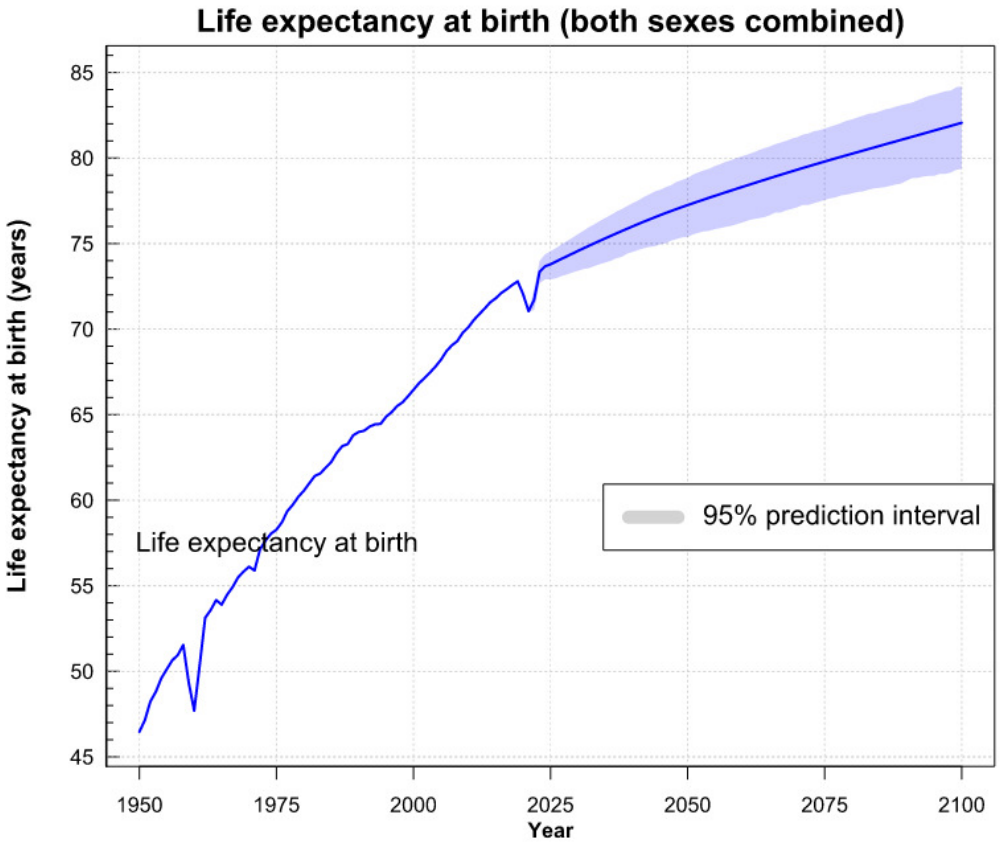
United Nations | Department of Economic and Social Affairs
Population Division



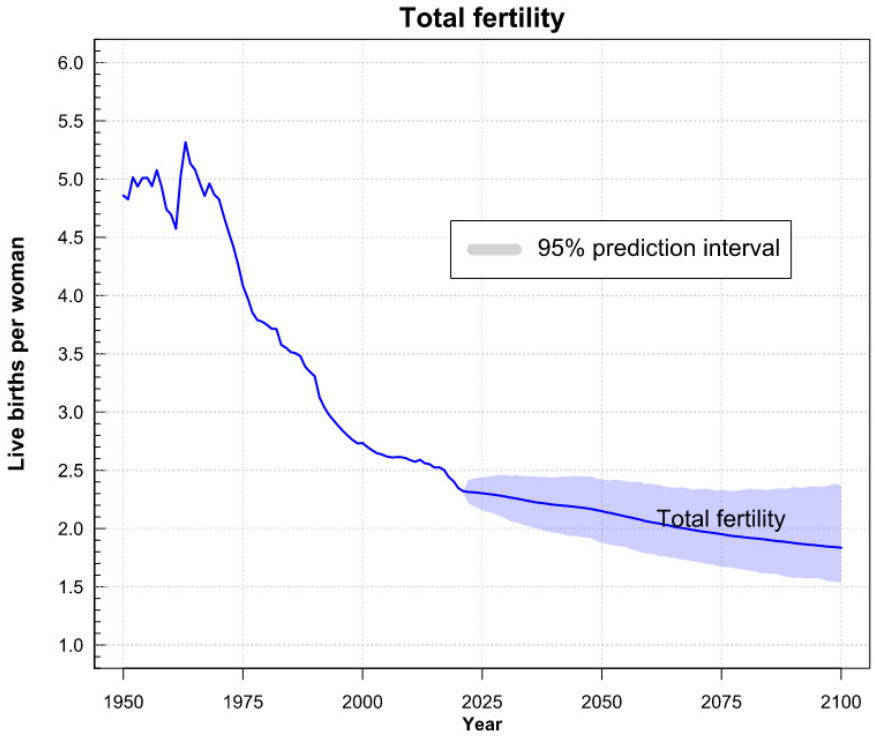
United Nations, DESA, Population Division. *World Population Prospects 2022*. http://population.un.org/wpp/

UN (2022). World Population Prospects. Web link.

Predictive modeling in demography: Projections and forecasts



United Nations | Department of Economic and Social Affairs
Population Division



UN (2022). World Population Prospects. Web link.

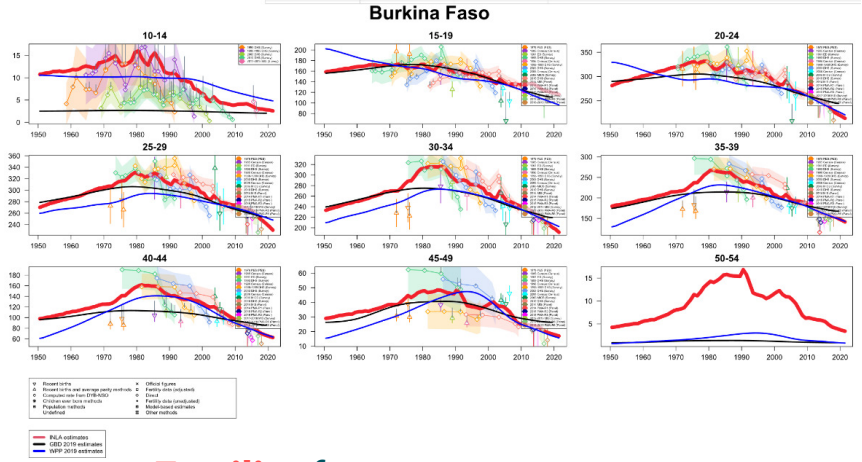
Predictive modeling in demography: Projections and forecasts

Population projection Cohort component / Leslie

$$\begin{pmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ \vdots \\ n_{\nu-1} \end{pmatrix}_{t+1} = \begin{pmatrix} f_0 & f_1 & f_2 & \dots & f_{\nu-2} & f_{\nu-1} \\ s_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & s_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & s_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & s_{\nu-2} & 0 \end{pmatrix} \cdot \begin{pmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ \vdots \\ n_{\nu-1} \end{pmatrix}_t$$

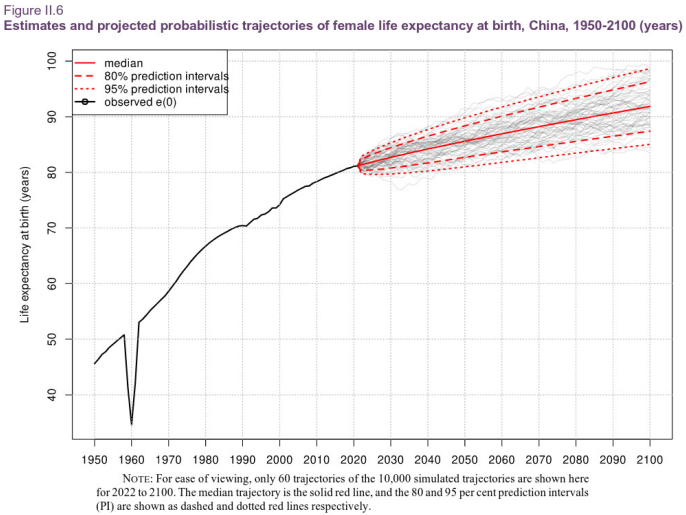
+ Net-migration by age

Baseline **population estimates**



Fertility forecasts

United Nations | Department of Economic and Social Affairs Population Division



Life-expectancy forecasts with derived mortality by age

UN (2022). World Population Prospects. Web link.

Predictive modeling in demography: Projections and forecasts

$$\begin{pmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ \vdots \\ n_{\nu-1} \end{pmatrix}_{t+1} = \begin{pmatrix} f_0 & f_1 & f_2 & \dots & f_{\nu-2} & f_{\nu-1} \\ s_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & s_1 & 0 & \dots & 0 & 0 \\ 0 & 0 & s_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & s_{\nu-2} & 0 \end{pmatrix} \cdot \begin{pmatrix} n_0 \\ n_1 \\ n_2 \\ n_3 \\ \vdots \\ n_{\nu-1} \end{pmatrix}_t$$

United Nations | Department of Economic and Social Affairs
Population Division

Figure II.6
Estimates and projected probabilistic trajectories of female life expectancy at birth, China, 1950-2100 (years)

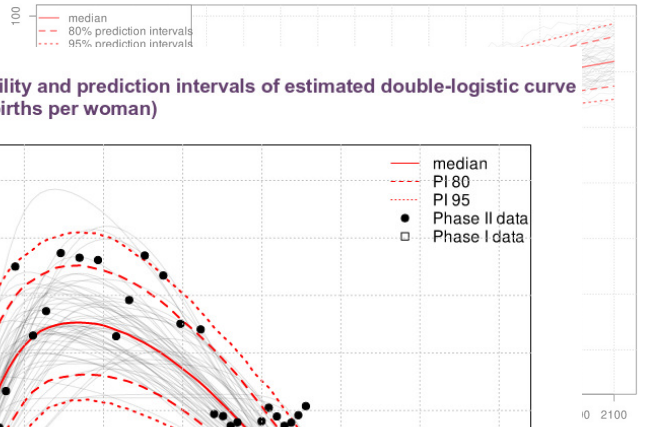
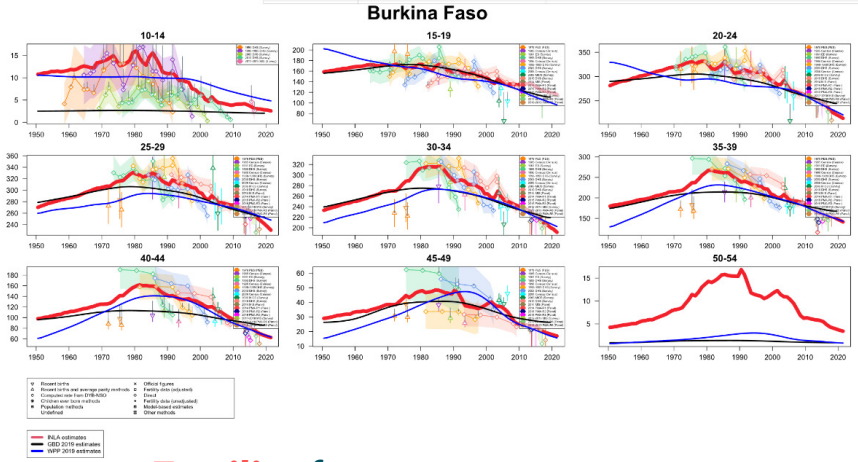


Figure II.2
Total fertility annual decrements by level of fertility and prediction intervals of estimated double-logistic curve for Bangladesh (systematic decline part) (live births per woman)

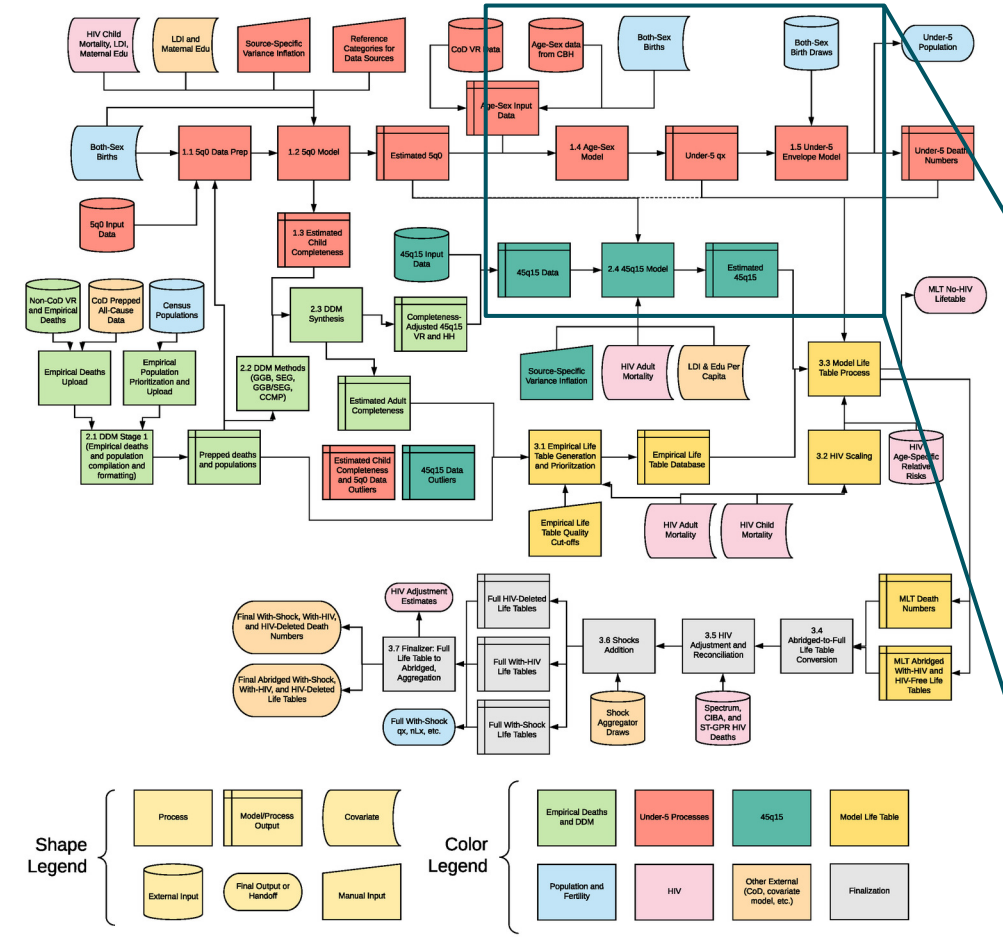
ith



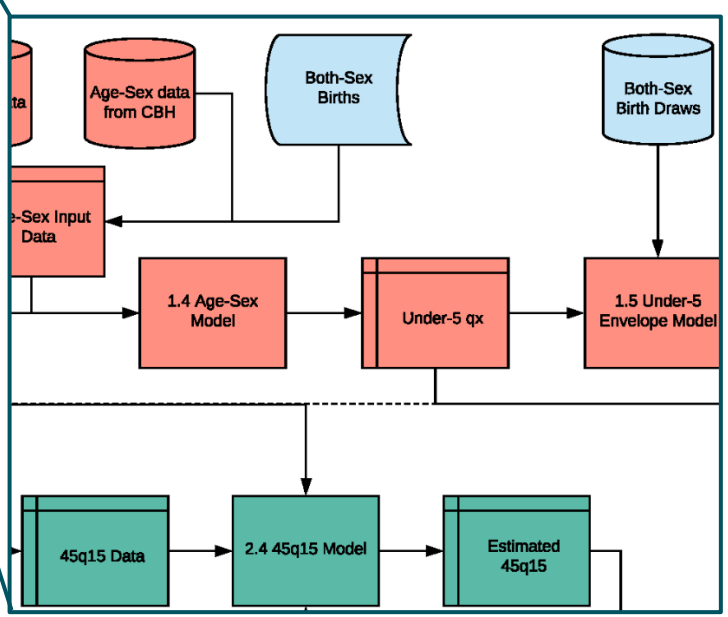
Fertility forecasts

UN (2022). World Population Prospects. Web link.

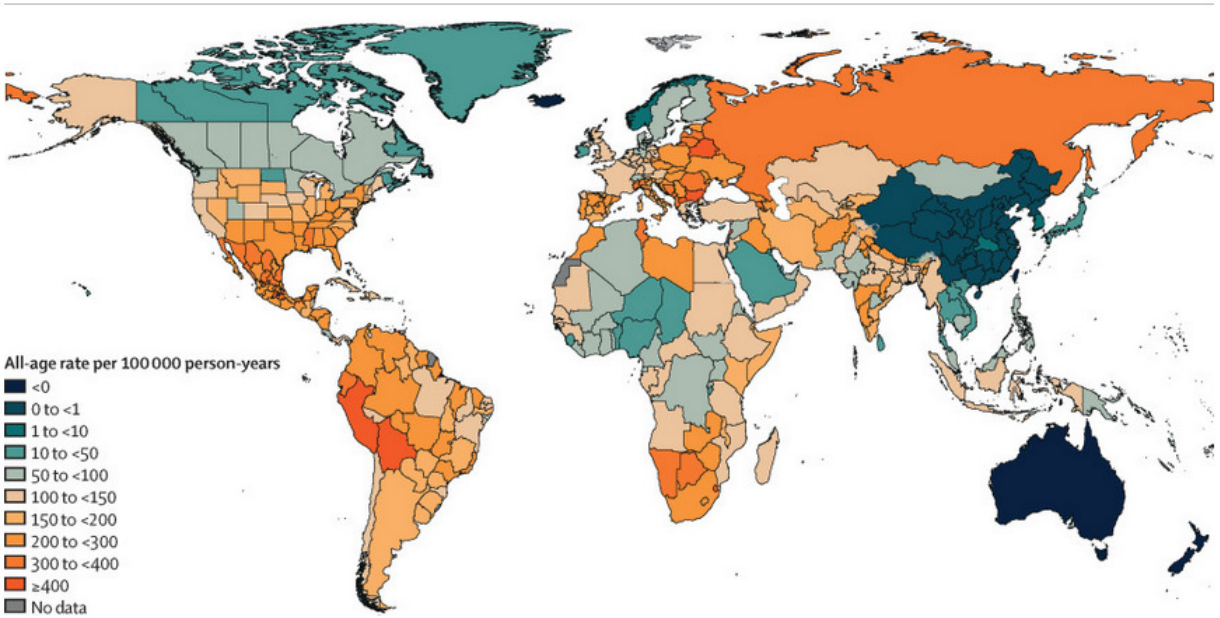
Predictive modeling in demography: Projections and forecasts



Global Burden of Disease (2019).
Mortality 1 Estimation Flowchart. Web link.



Predictive modeling in demography: Indirect estimation



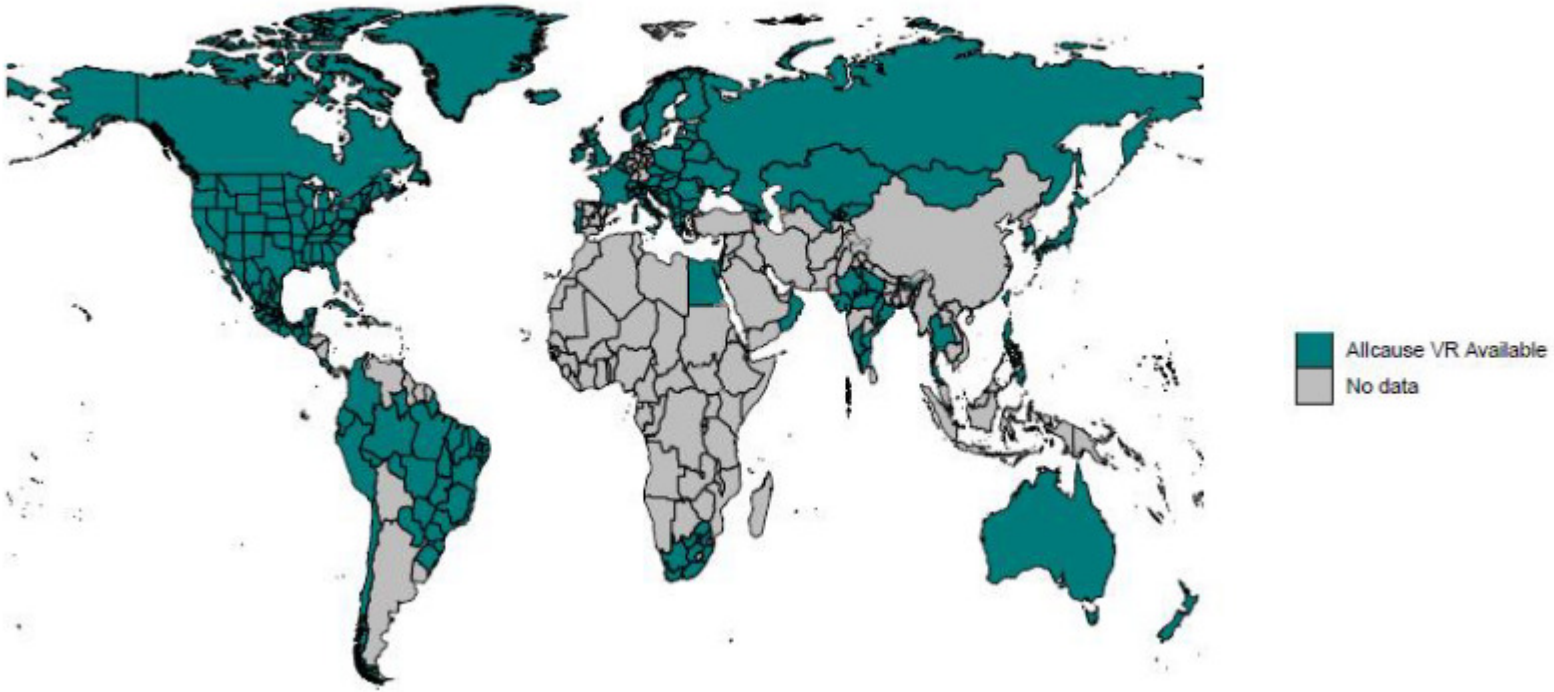
Download : Download high-res image (1MB) Download : Download full-size image

Figure 2. Global distribution of estimated excess mortality rate due to the COVID-19 pandemic, for the cumulative period 2020–21

COVID-19 Excess Mortality Collaborators (2022). Estimating excess mortality... 10.1016/S0140-6736(21)02796-3

Predictive modeling in demography: Indirect estimation

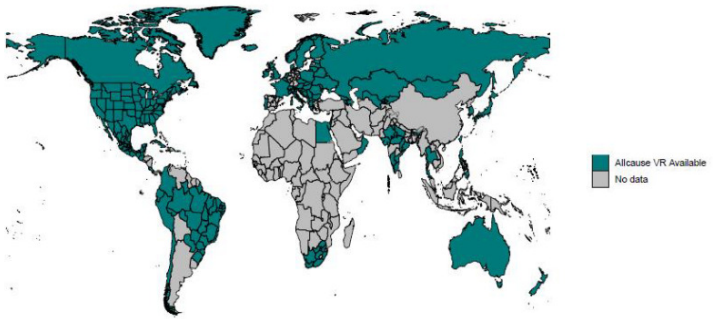
Figure S1. Map of all cause data availability



COVID-19 Excess Mortality Collaborators (2022). Estimating excess mortality... Link to Supplementary

Predictive modeling in demography: Indirect estimation

Figure S1. Map of all cause data availability



Download : Download high-res image (1MB) Download : Download full-size image

Figure 2. Global distribution of estimated excess mortality rate due to the COVID-19 pandemic, for the cumulative period 2020–21

To arrive at a parsimonious model, we used the Least Absolute Shrinkage and Selection Operator (LASSO) regression to help select a list of covariates that have sensible direction of effect on the excess mortality rate due to the COVID-19 pandemic. The final model can be described by the equation below:

$$\ln(y_i) = \alpha + \sum_{j=1}^{16} \beta_j \cdot x_{ij} + \epsilon_i$$

Average absolute latitude		Positive
Cardiovascular diseases death rate (2019)	Log	Positive
Crude death rate (2019)	Log	Positive
Diabetes death rate (2019)	Log	Positive
Healthcare access and quality Index (2019)		Negative
HIV death rate (2019)	Log	Positive
Infection detection ratio (lagged)		Negative
Inpatient admission rate (2019)		Negative
Mobility (lagged)		Positive
Proportion of population over age 75		Positive
Quality of death registration system (2019)		Negative
Reported COVID-19 death rate	Log	Positive
Seroprevalence rate (lagged)	Log	Positive
Smoking prevalence (2019)		Positive
Universal health coverage (2019)		Negative

COVID-19 Excess Mortality Collaborators (2022). Estimating excess mortality... Link to Supplementary

github.com/ihmeuw-demographics/publication_covid_em/tree/main/03_ensemble_excess_model

Parameter-centric modeling vs. predictive modeling

$$f(y|x,\theta)$$

vs.

$$f(y|x,\theta)$$

Predictive modeling workflow

$$f(y|x,\theta)$$

Predict from model
Transform model predictions
Propagate prediction uncertainty through analysis
Cross-validate predictive accuracy & model calibration

Predict from model

Predict from model – The model equation

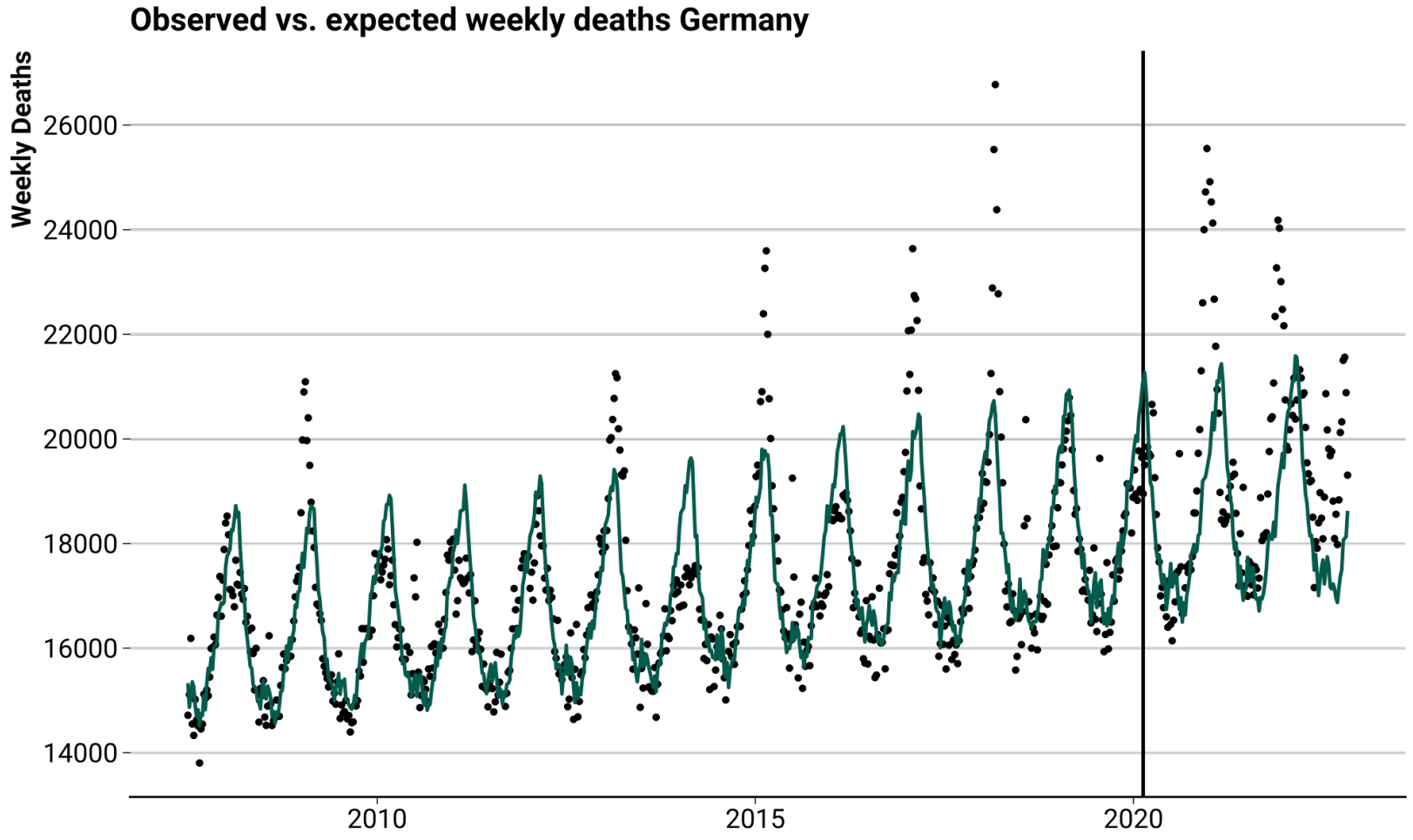
$$D_t \sim \text{NegBin}(\mu_t, \theta)$$

$$\mu_t = \exp(\beta_0 + \beta_y \text{year}_t + \Sigma_{i=1:52} \beta_{w[i]} \text{week}_t + \log(\text{population}))$$

$$\theta = \exp(\beta_\theta)$$

Observations
Sampling Distribution
Parameters
Predictors

Predict from model – Point estimates



Transform model predictions

Transform model predictions

$$D_t \sim \text{NegBin}(\mu_t, \theta)$$

$$\mu_t = \exp(\beta_0 + \beta_y \text{year}_t + \Sigma_{i=1:52} \beta_{w[i]} \text{week}_t + \log(\text{population}))$$

$$\theta = \exp(\beta_\theta)$$

Cumulative excess deaths:

$$\Sigma_t D_t - \mu_t$$

Transform model predictions

Propagate prediction uncertainty through analysis

Propagate uncertainty

How many deaths in a week had C19 not occurred?

Excess = Observed - Expected

Propagate uncertainty

How many deaths in a week had C19 not occurred?

Excess = Observed - Expected

Propagate uncertainty

How many deaths in a week had C19 not occurred?

Excess = Observed - Expected

Propagate uncertainty – Epistemic uncertainty

$$D_t \sim \text{NegBin}(\mu_t, \theta)$$

$$\mu_t = \exp(\beta_0 + \beta_y \text{year}_t + \Sigma_{i=1:52} \beta_{w[i]} \text{week}_t + \log(\text{population}))$$

$$\theta = \exp(\beta_\theta)$$

Uncertainty regarding the state of the world, here:
Uncertain parameter estimates

```
Family: Negative Binomial(404.799)
Link function: log

Formula:
deaths_observed ~ iso_year + iso_week_fct + offset(log(personweeks))

Parametric coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.427e+01  1.102e+00 -22.028  < 2e-16 ***
iso_year        7.886e-03  5.471e-04  14.413  < 2e-16 ***
iso_week_fct2   1.002e-02  2.051e-02   0.489  0.62508
iso_week_fct3   7.545e-03  2.051e-02   0.368  0.71298
iso_week_fct4   1.041e-02  2.051e-02   0.508  0.61167
iso_week_fct5   3.438e-02  2.051e-02   1.677  0.09364 .
iso_week_fct6   3.924e-02  2.051e-02   1.913  0.05570 .
iso_week_fct7   5.035e-02  2.051e-02   2.455  0.01408 *
iso_week_fct8   5.005e-02  2.051e-02   2.440  0.00398 **
```

Hüllermeier (2021). Aleatoric and epistemic uncertainty in machine learning... 10.1007/s10994-021-05946-3

Propagate uncertainty – Epistemic uncertainty

7.5.2 The delta method

The delta method provides an easy approximation for the confidence limits on values that are not parameters of the model. To use it you must have a formula for $\mu = f(a, b)$ that you can differentiate with respect to a and b. Unlike the first likelihood profile method, you don't have to be able to solve the equation for one of the parameters.

The formula for the delta method comes from a Taylor expansion of the formula for μ, combined with the definitions of the variance ($V(a) = E[(a - \bar{a})^2]$) and covariance ($C(a, b) = E[(a - \bar{a})(b - \bar{b})]$):

$$V(f(a,b)) \approx V(a)\left(\frac{\partial f}{\partial a}\right)^2 + V(b)\left(\frac{\partial f}{\partial b}\right)^2 + 2C(a,b)\frac{\partial f}{\partial a}\frac{\partial f}{\partial b}. \qquad (7.5.4)$$

See the Appendix, or Lyons (1991) for a derivation and details.

7.5.3 Population prediction intervals (PPI)

Another simple procedure for calculating confidence limits is to draw random samples from the estimated sampling distribution (approximated by the information matrix) of the parameters. In the approximate limit where the information matrix approach is valid, it turns out that the distribution of the parameters will be multivariate normal with a variance-covariance matrix given by the inverse of the information matrix. The `MASS` package in R has a function, `mvrnorm`*, for selecting multivariate normal random deviates. With the `mle2` fit `w1` from above, then

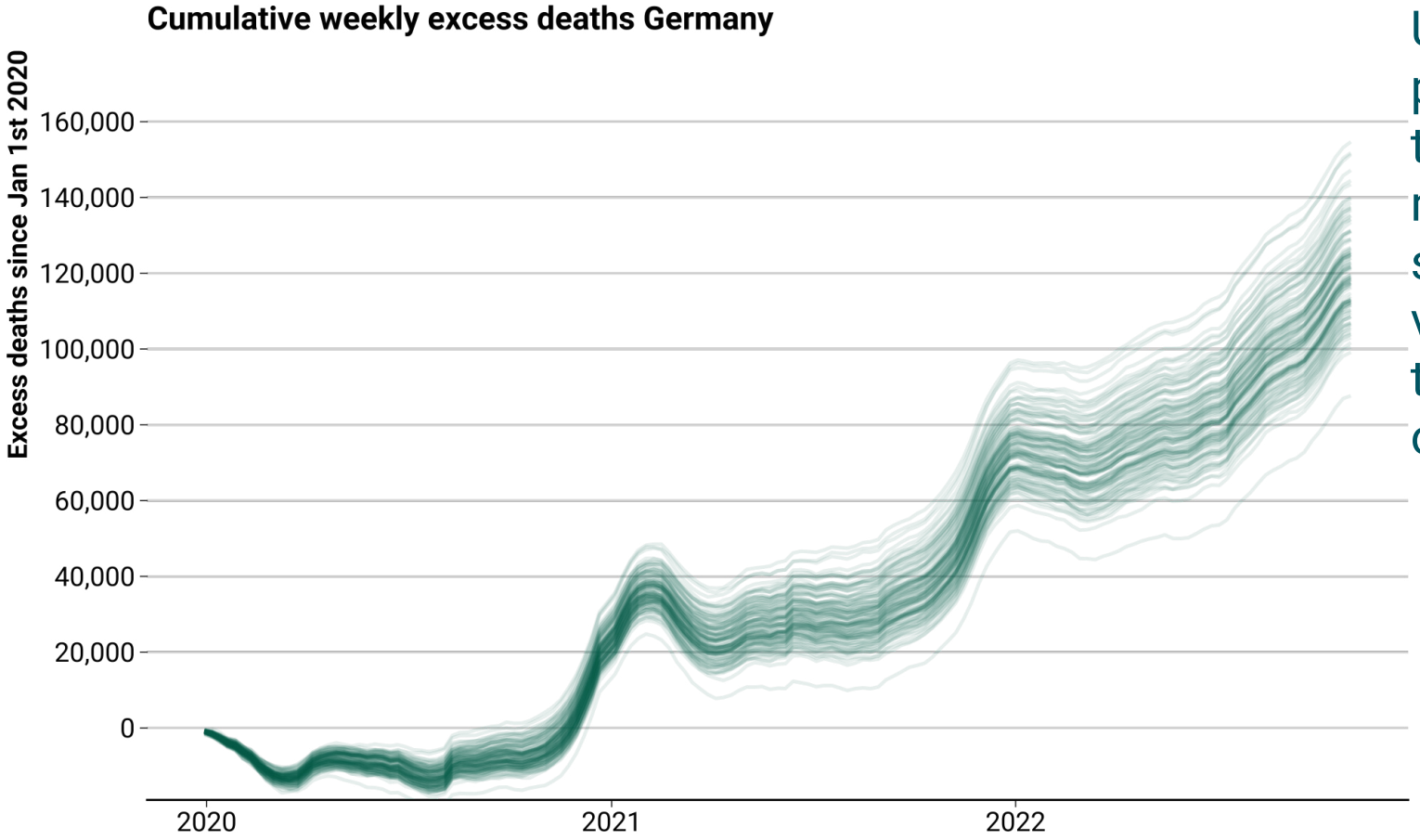
```
> vmat = mvrnorm(1000, mu = coef(w1), Sigma = vcov(w1))
```

7.5.4 Bayesian analysis

Finally, you can use a real Bayesian method: construct either an exact Bayesian model, or, more likely, a Markov chain Monte Carlo analysis for the parameters. Then you can calculate the posterior distribution of any function of the parameters (such as the mean survival time) from the posterior samples of the parameters, and get the 95% credible interval.

Bolker (2007). Ecological Models and Data in R. ms.mcmaster.ca/~bolker/emdbook/book.pdf

Propagate uncertainty – Epistemic uncertainty



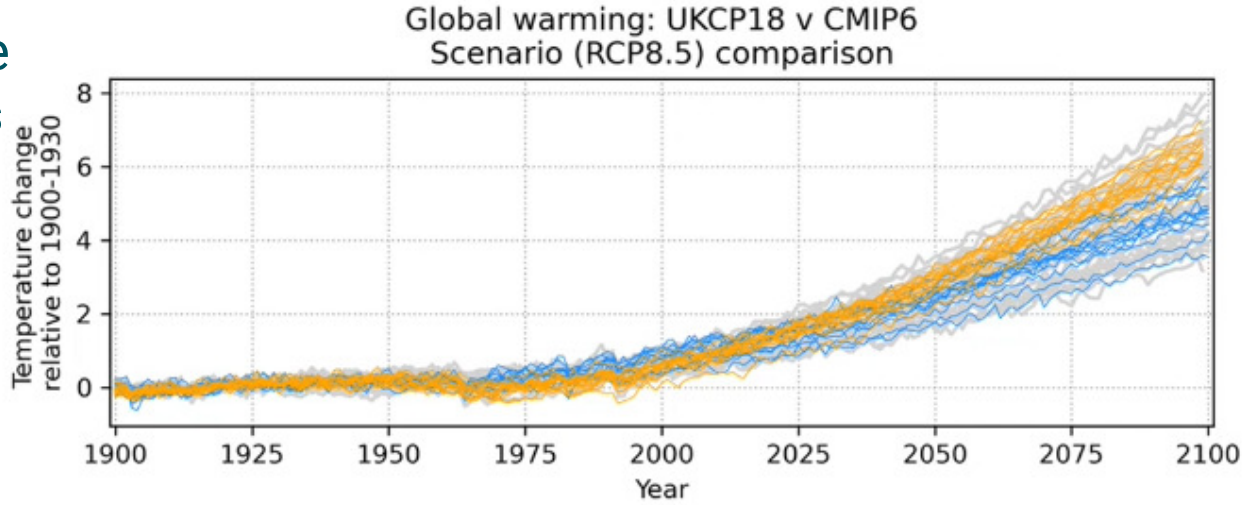
Uncertainty in model parameters propagated through transformed model predictions via sampling from Multi-variate Normal approx. to sampling distribution of beta's

Propagate uncertainty – Epistemic uncertainty

Uncertainty regarding the state of the world, here:
Uncertain model structure

Uncertainties in projected changes

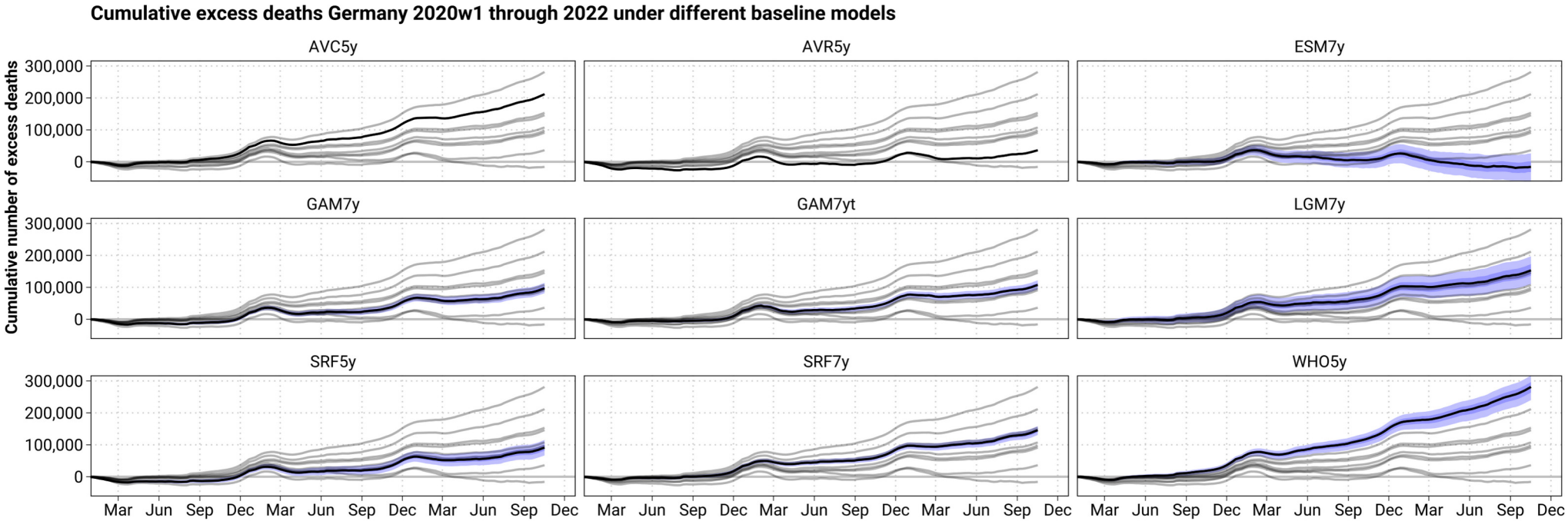
Internal variability and modelling uncertainty



"The reason why different climate models give **different projections** is because they use **different, but plausible, representations** of climate processes."

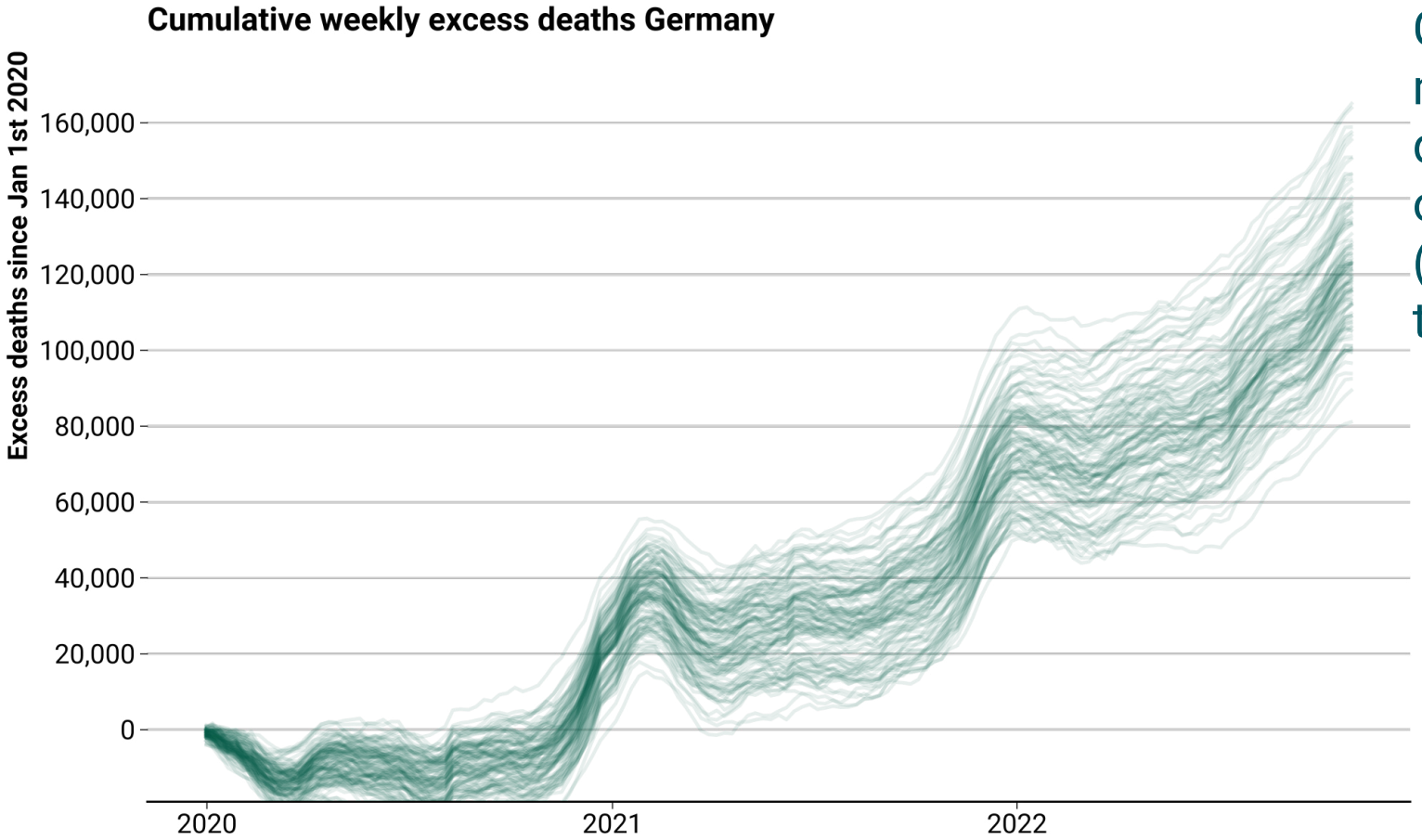
Murphy (2022). Informing risks and decisions using the UK climate projections. youtu.be/uiSTmy1vqIs

Propagate uncertainty – Epistemic uncertainty



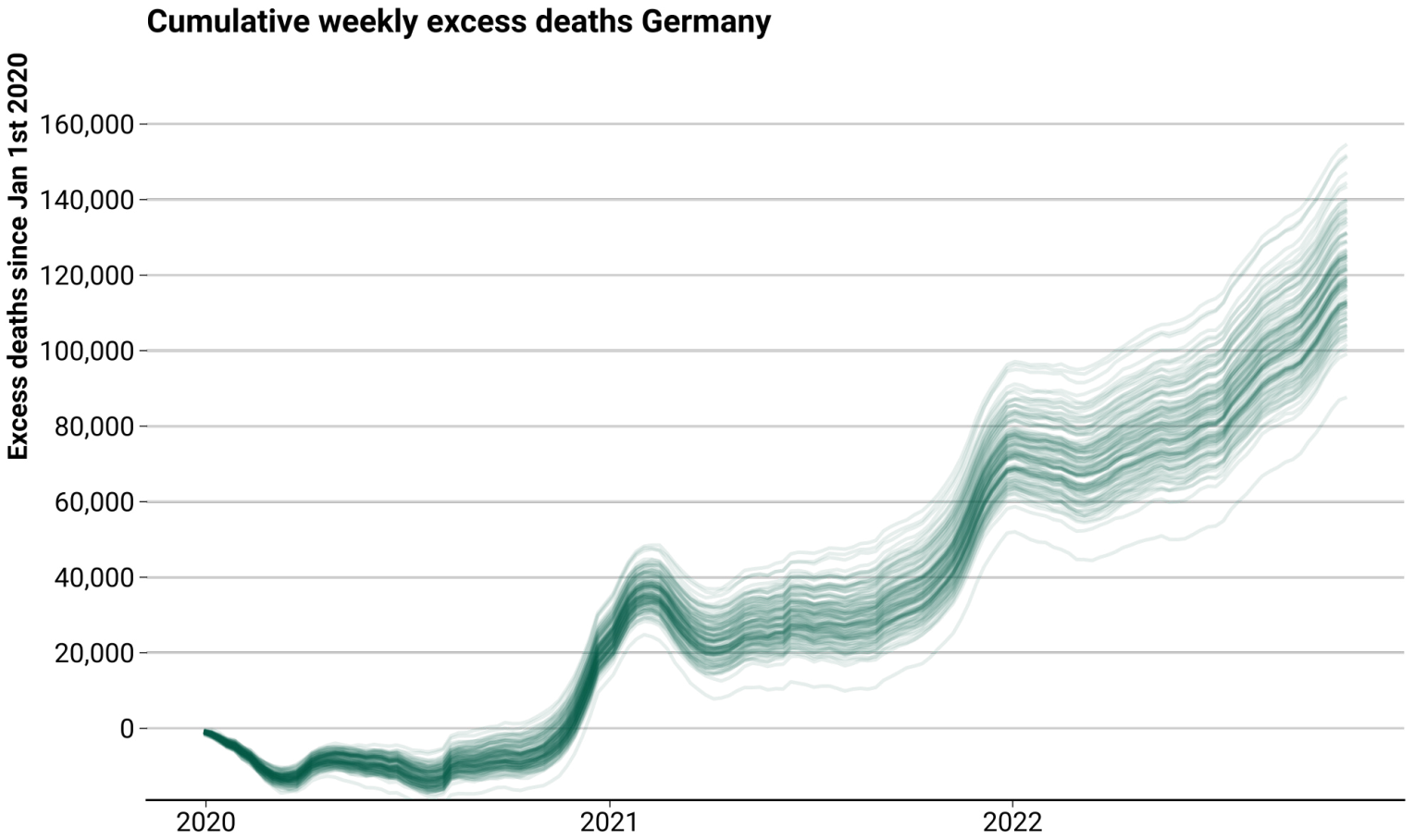
95 and 50% prediction intervals
AVC: 5 year avg. weekly death counts; AVR: 5 year avg. weekly death rates
GAM: Negative-binomial regression with smooth seasonality, control for mortality trends, population structure, public holidays; GAMt same as GAM with control for temperature anomalies
LGM: Quasi-Poisson regression with smooth seasonality, autocorrelation, control for mortality trends and population structure
SRF: Quasi-Poisson regression in Serfling specification trained on all weeks with control for population structure; WHO: WHO expected deaths specification
ESM: Seasonal exponential smoothing of log deaths
Jonas Schöley, PhD. Data: STMF mortality.org Nov 14 2022, Code: github.com/jschoeley/xgermany2020

Propagate uncertainty – Aleatoric uncertainty



Given the sampled model parameters, draw samples from the outcome distribution (Negative-Binomial in this case).

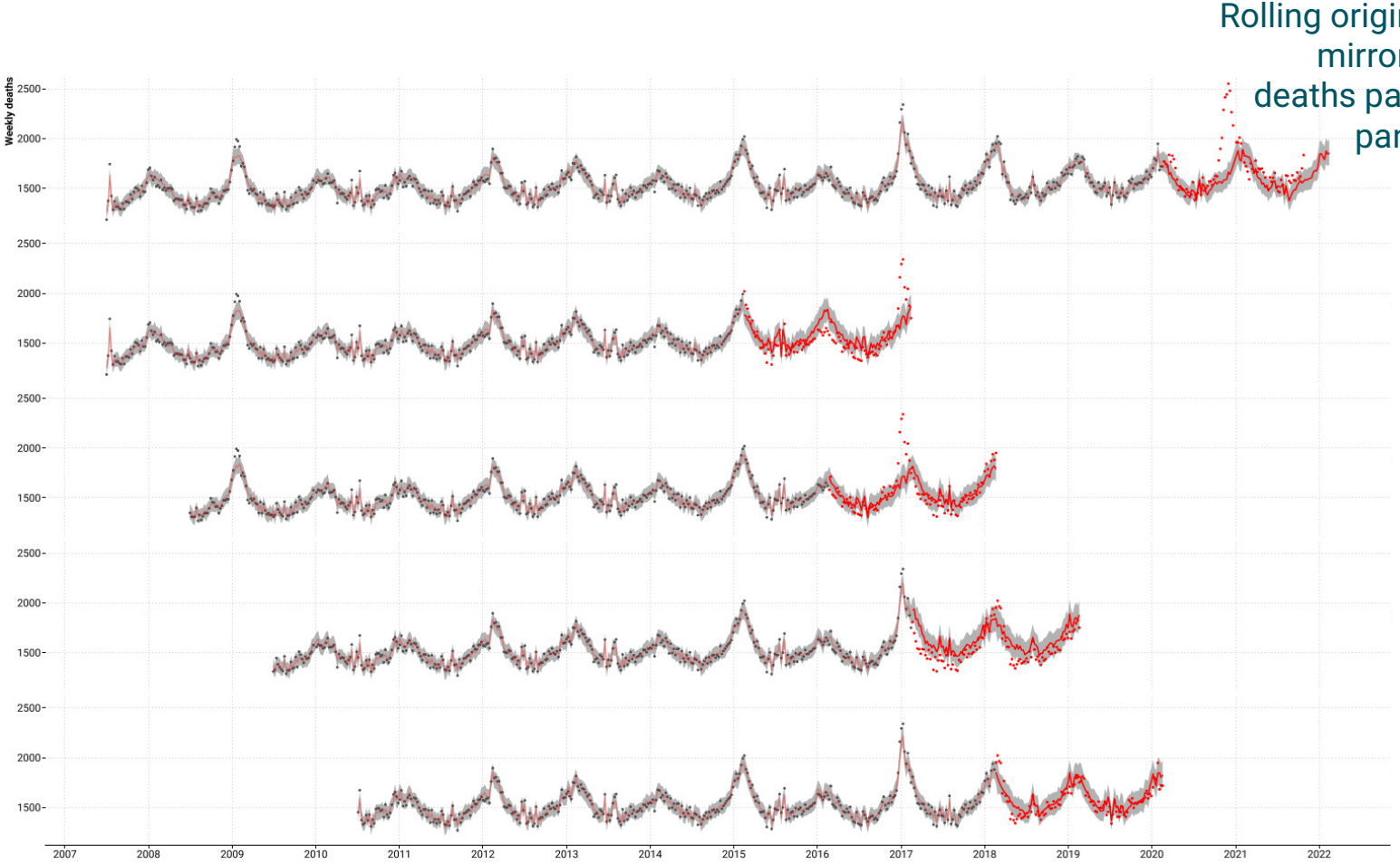
Propagate uncertainty – Aleatoric uncertainty



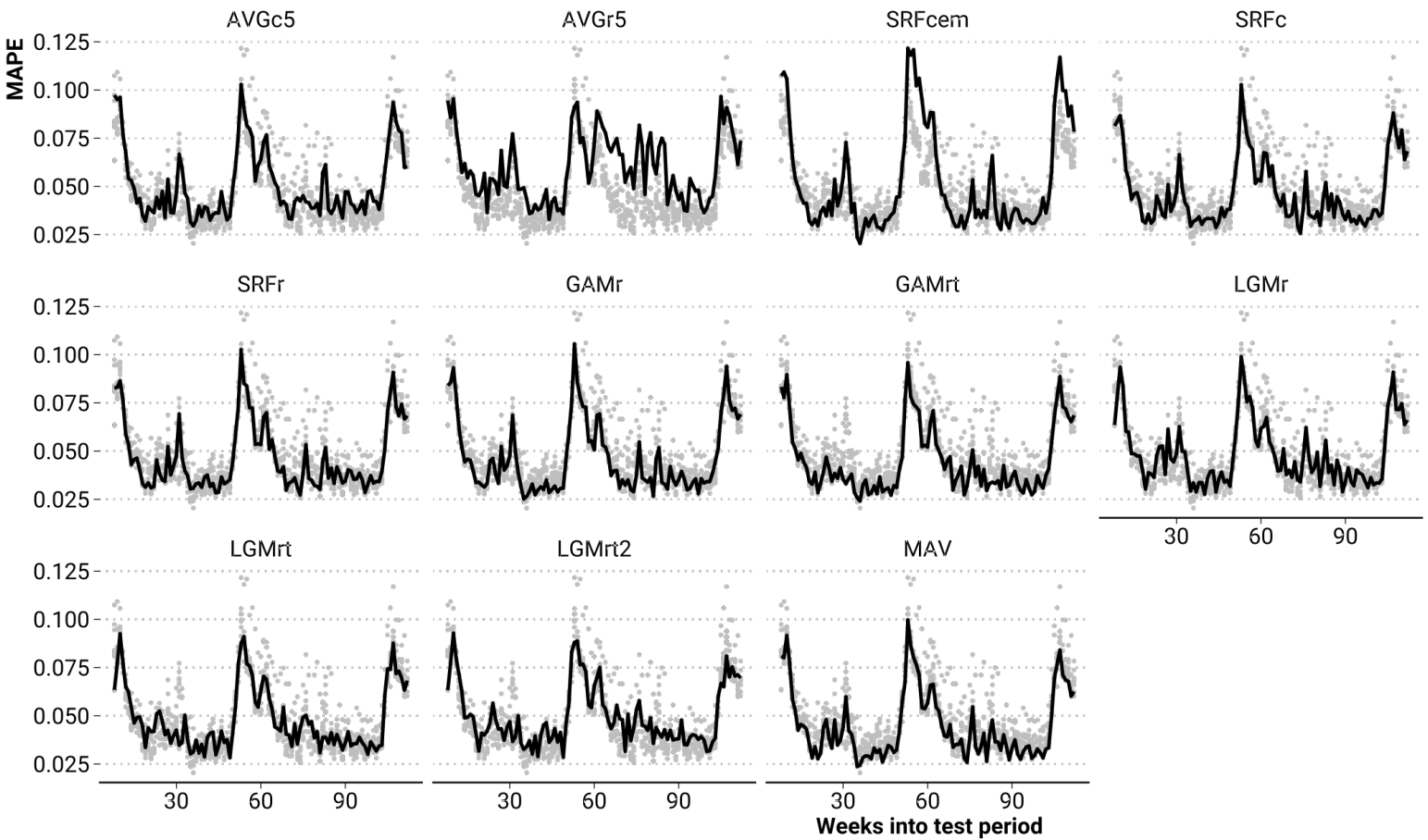
Cross-validate predictive accuracy & model calibration

Cross validate – Error & Bias

Rolling origin four-fold cross-validation setup mirroring the task of predicting weekly deaths past the beginning of the COVID-19 pandemic given pre-pandemic data.

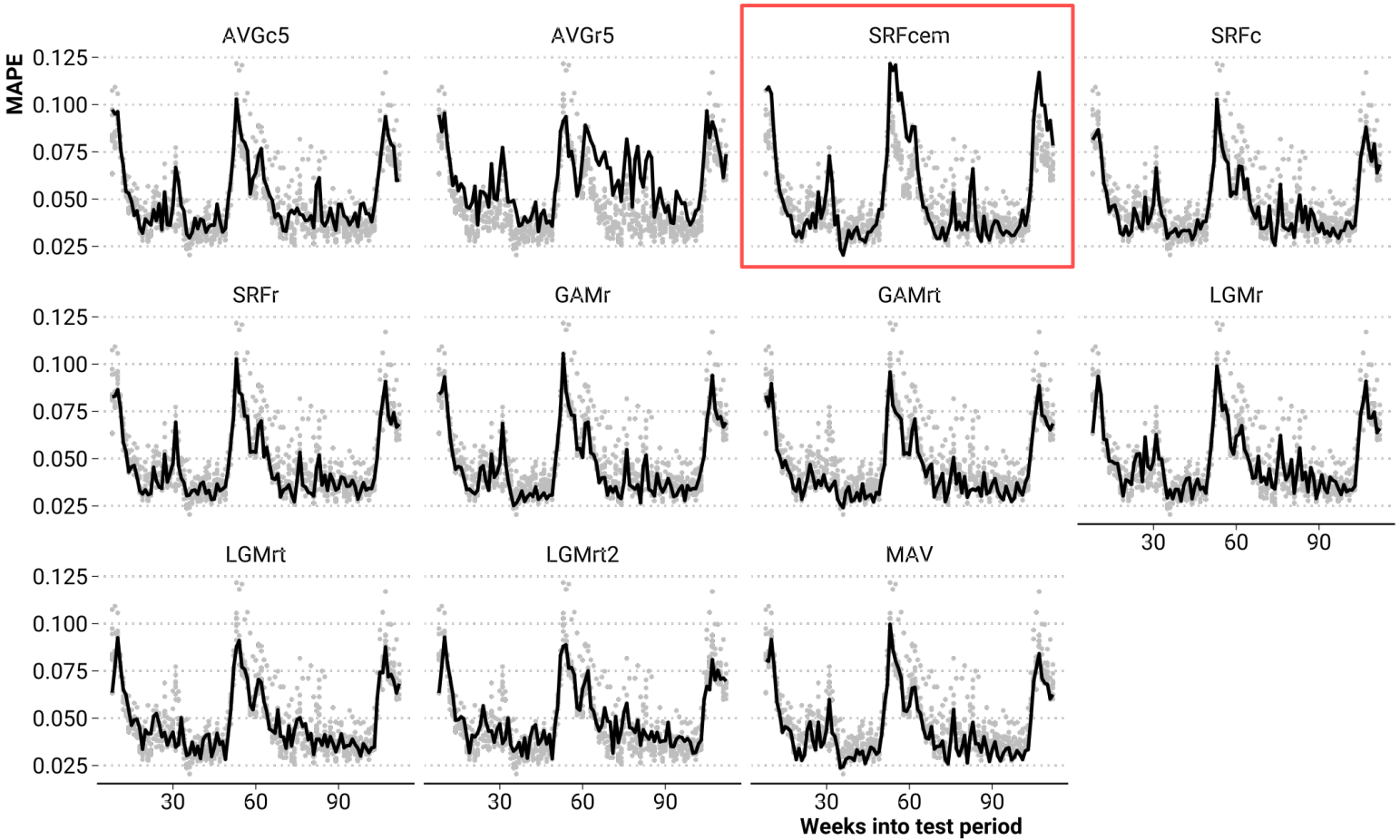


Cross validate – Error & Bias



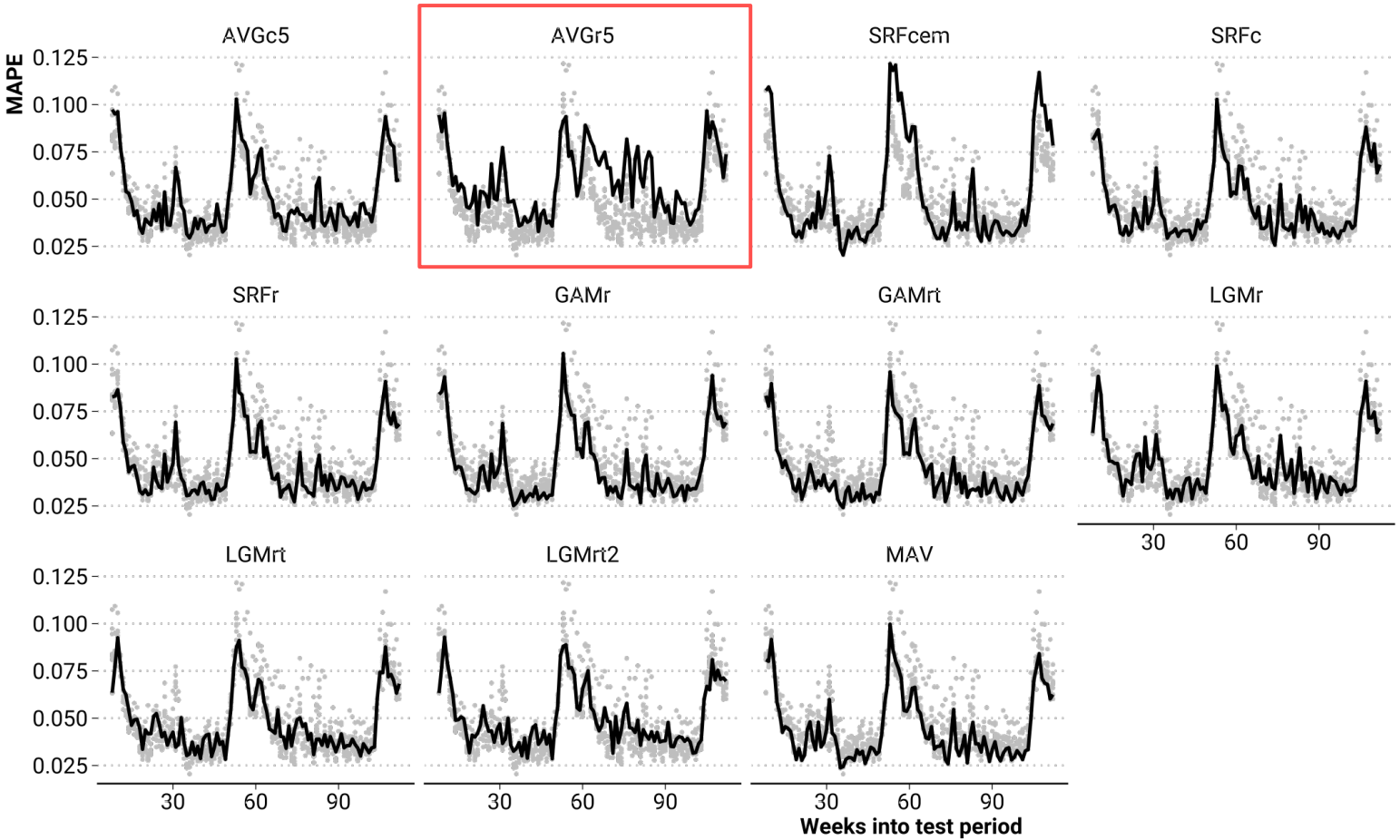
Mean absolute percentage error of weekly death counts for different models. Grey points show MAPE of all other models

Cross validate – Error & Bias



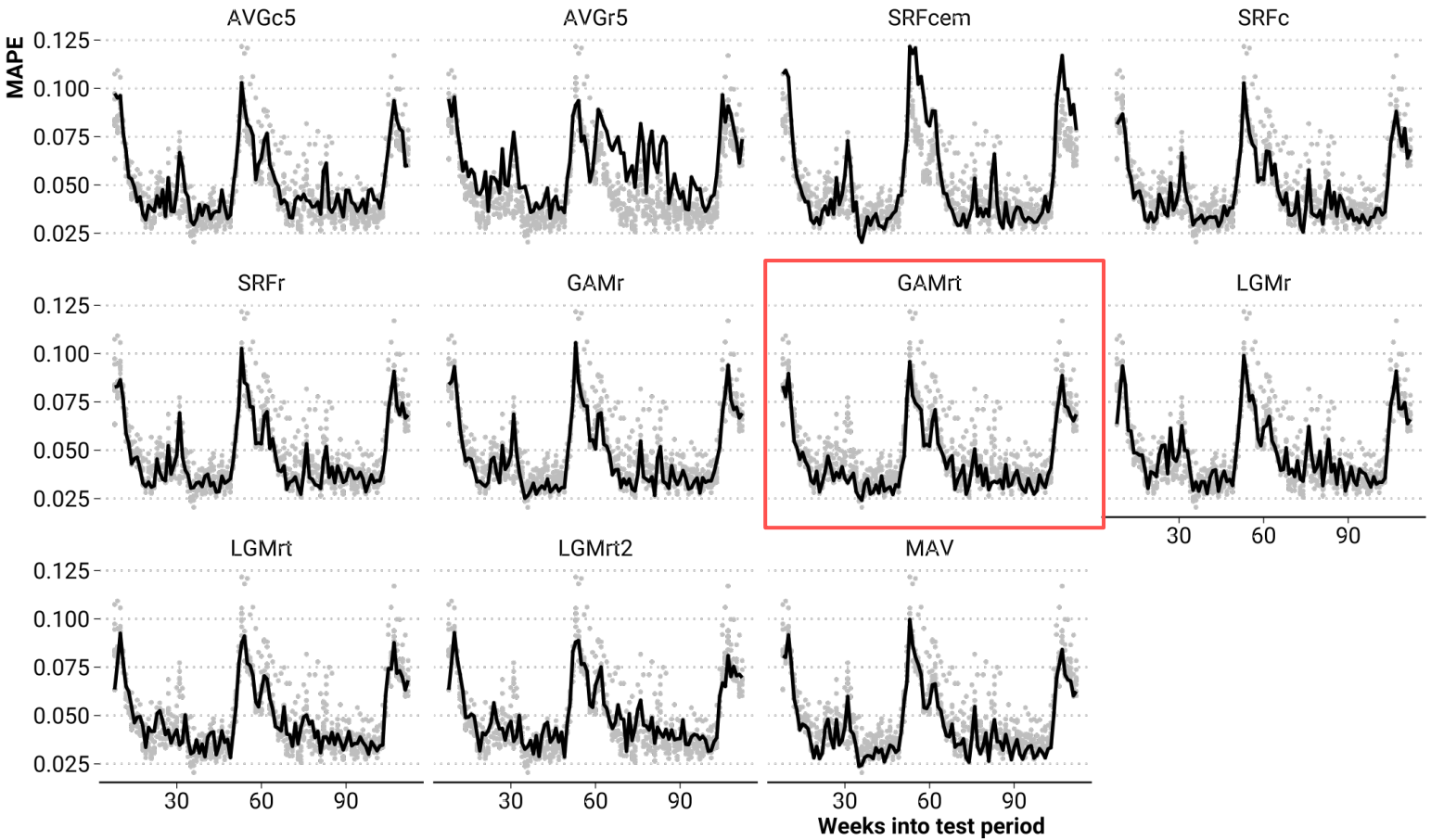
Mean absolute percentage error of weekly death counts for different models. Grey points show MAPE of all other models

Cross validate – Error & Bias



Mean absolute percentage error of weekly death counts for different models. Grey points show MAPE of all other models

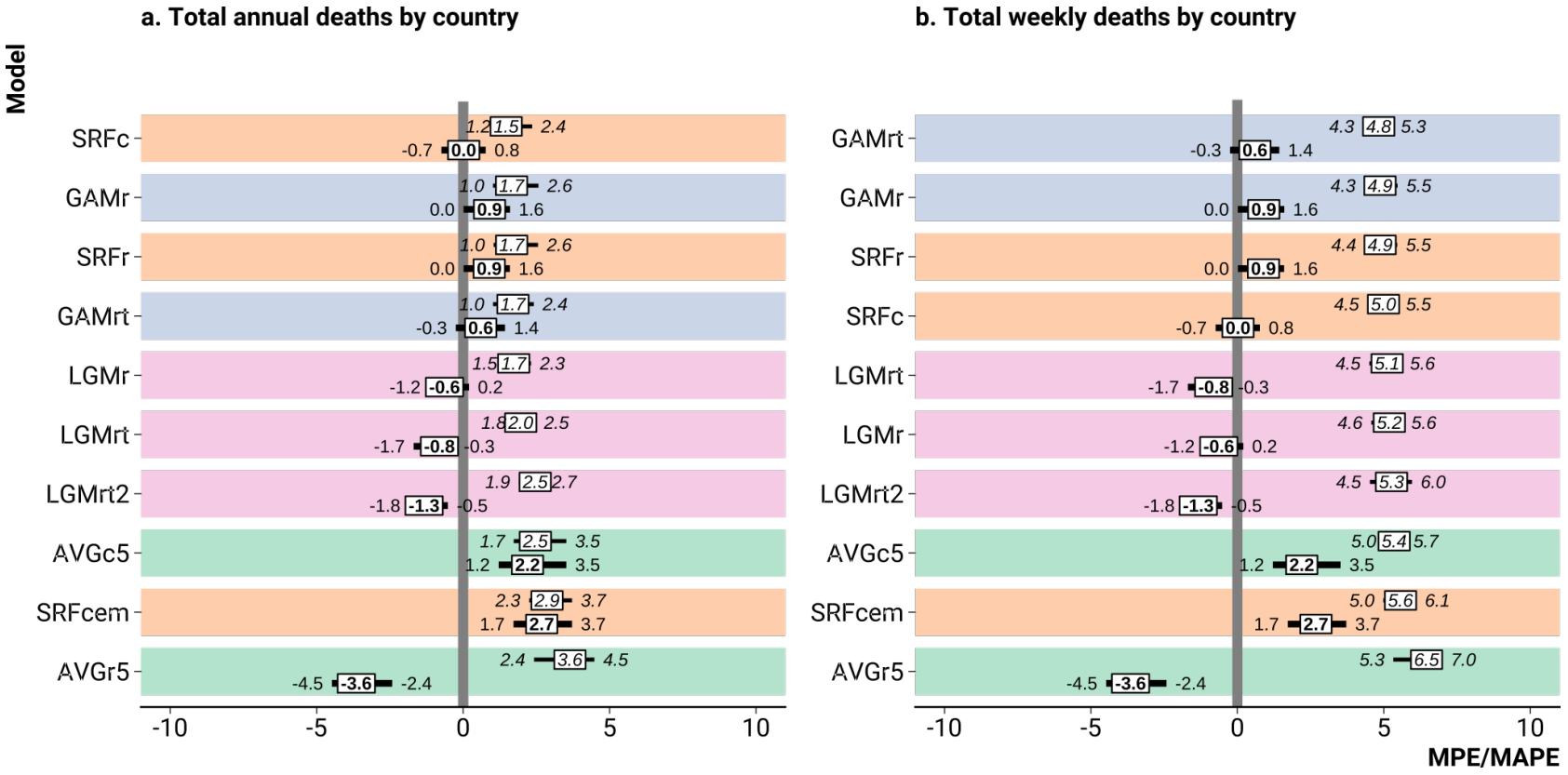
Cross validate – Error & Bias



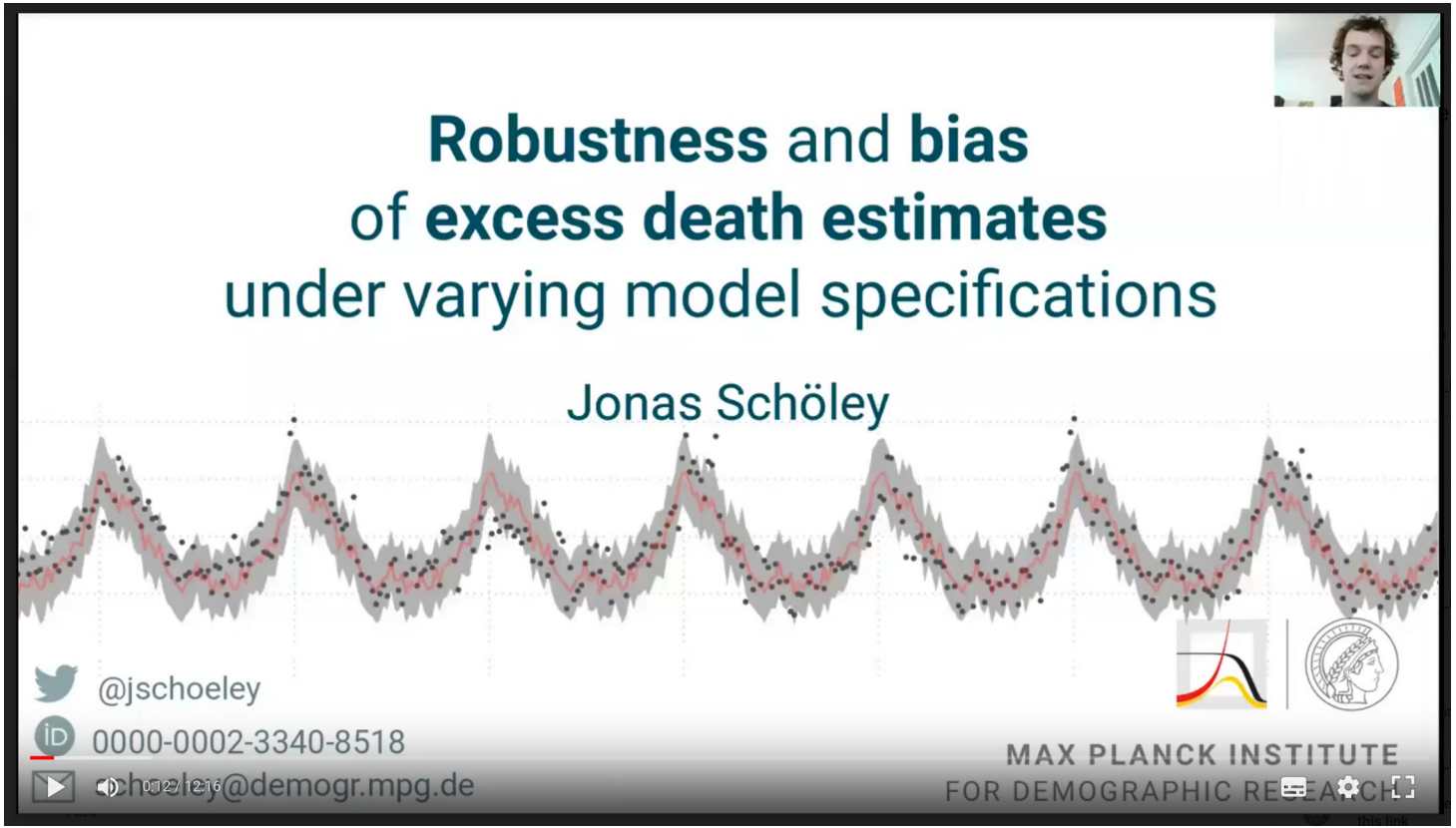
Mean absolute percentage error of weekly death counts for different models. Grey points show MAPE of all other models

Cross validate – Error & Bias

Bias (MPE, bold) and error (MAPE, italic) by model when predicting death counts on test data.

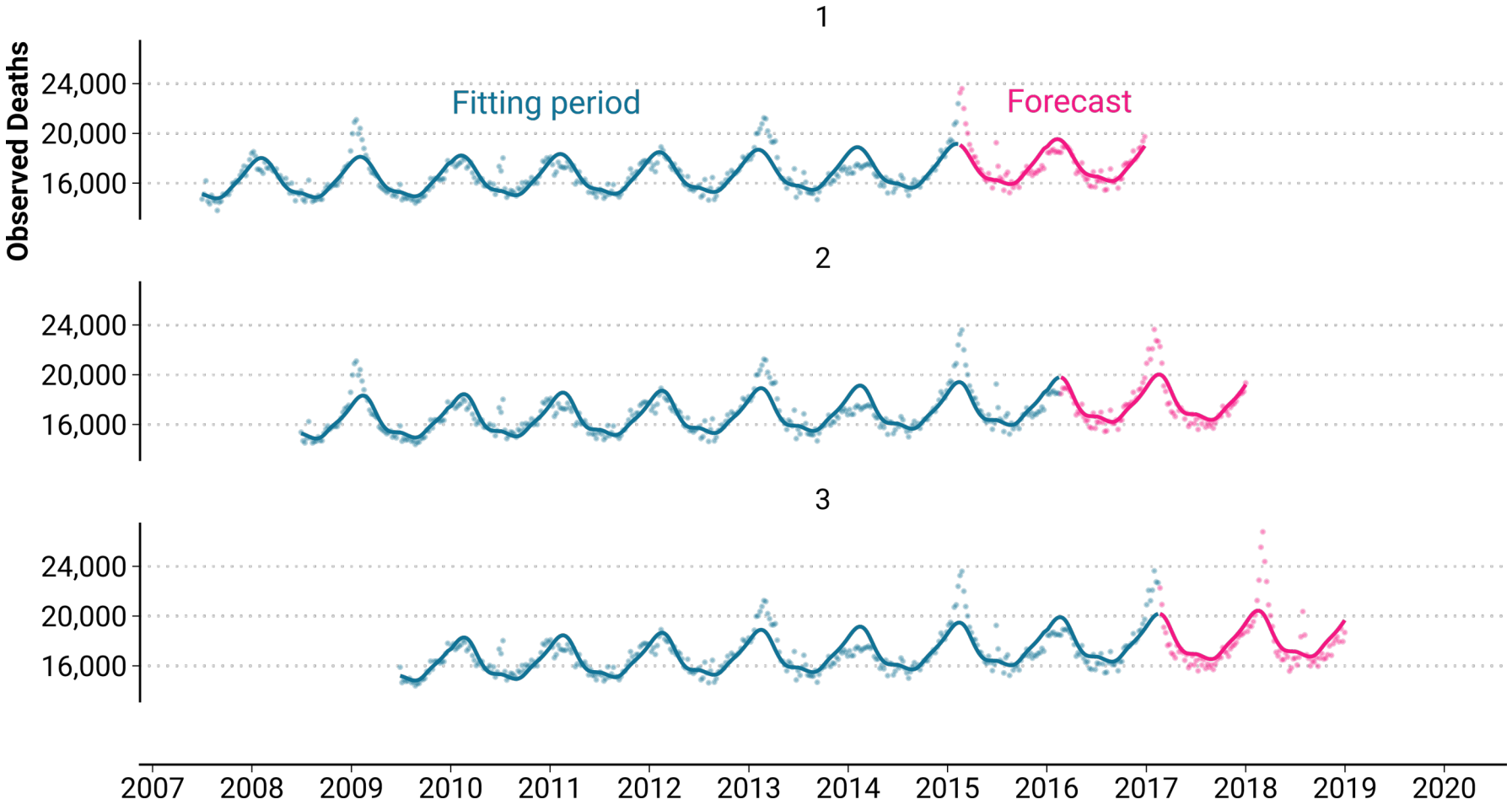


Cross validation – Error & Bias

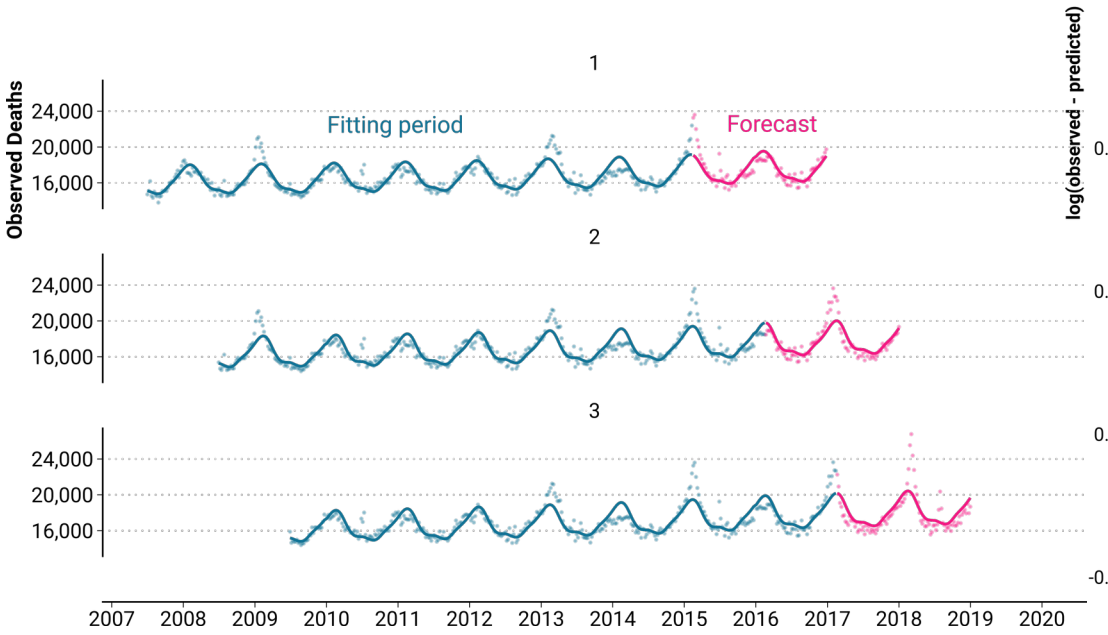


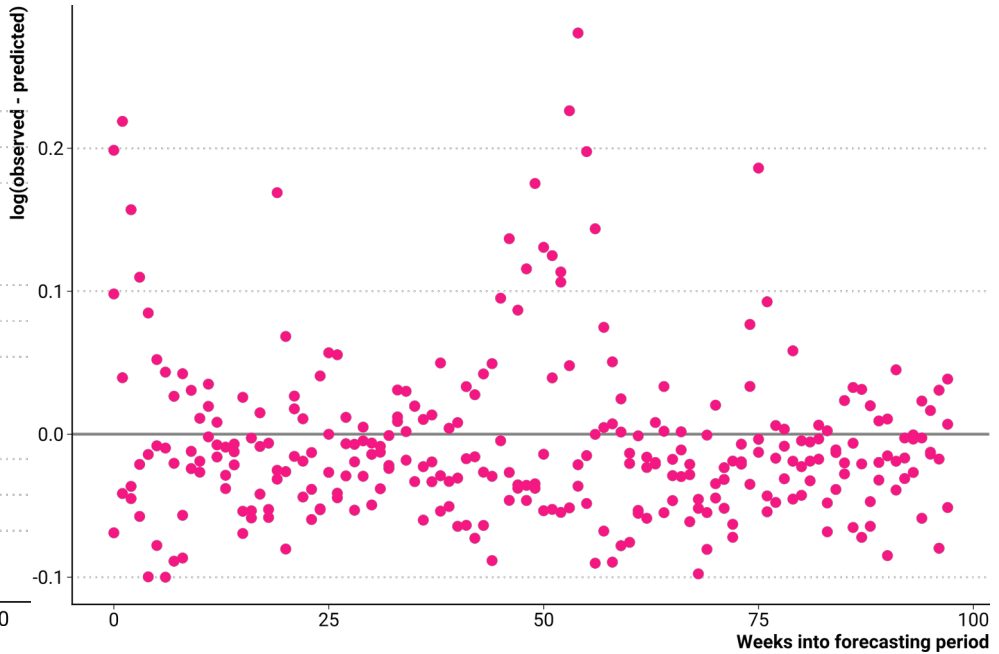
Schöley (2021). Robustness and bias in excess death estimates. Link to video

Cross validation – Calibration

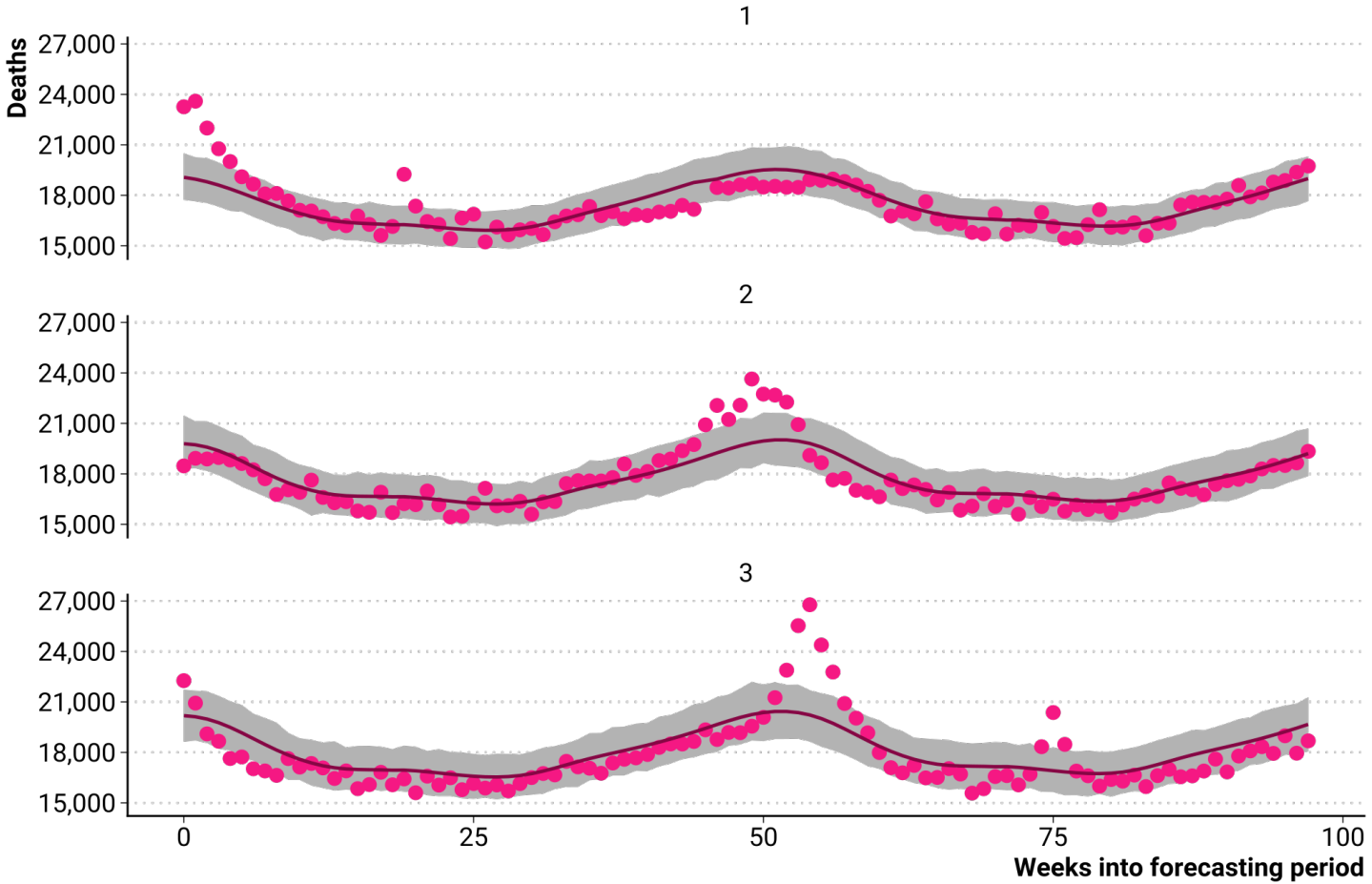


Cross validation – Calibration



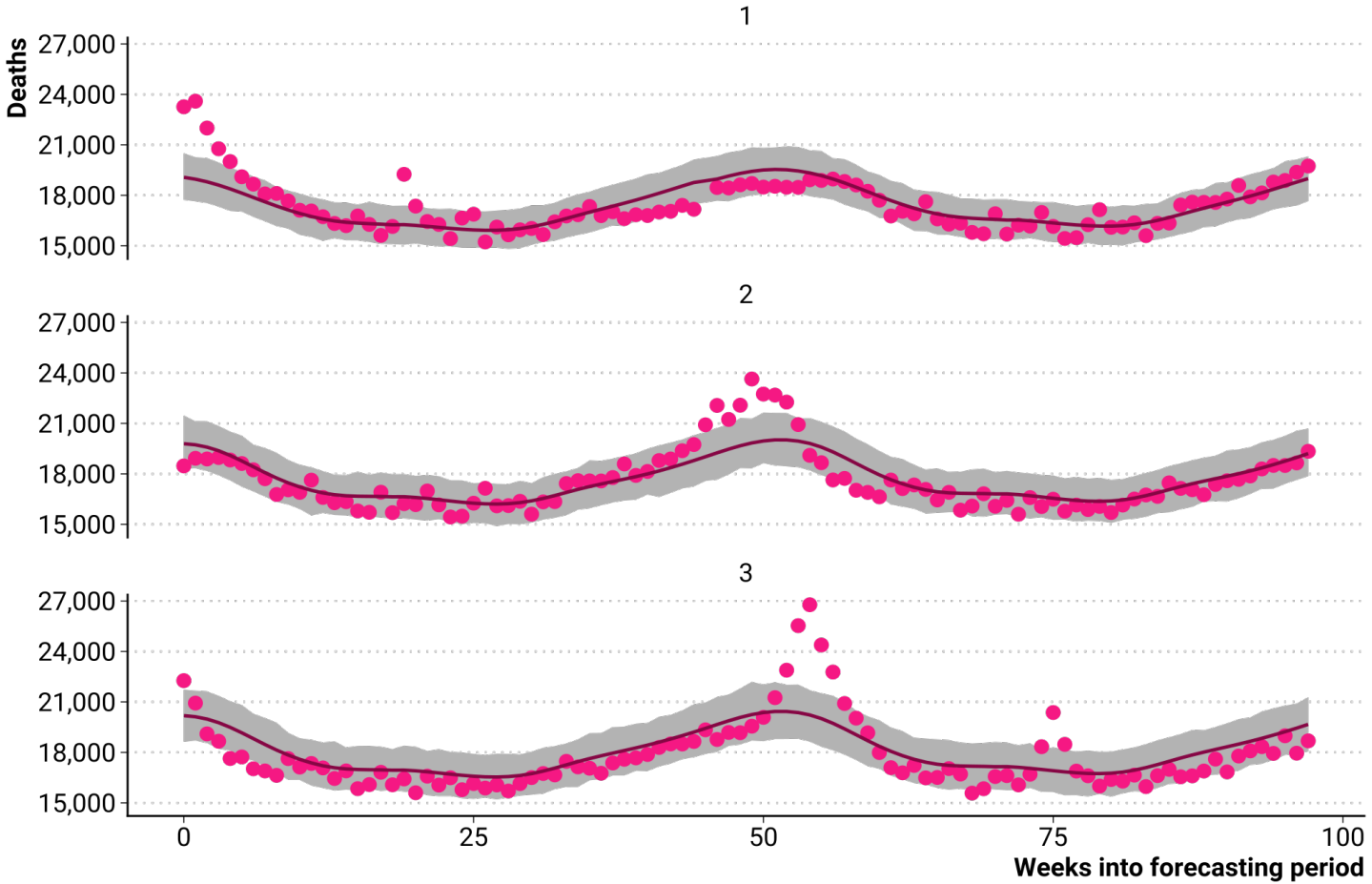


Cross validation – Calibration



Prediction interval with **90% nominal coverage** via Negative-Bin. GAM model

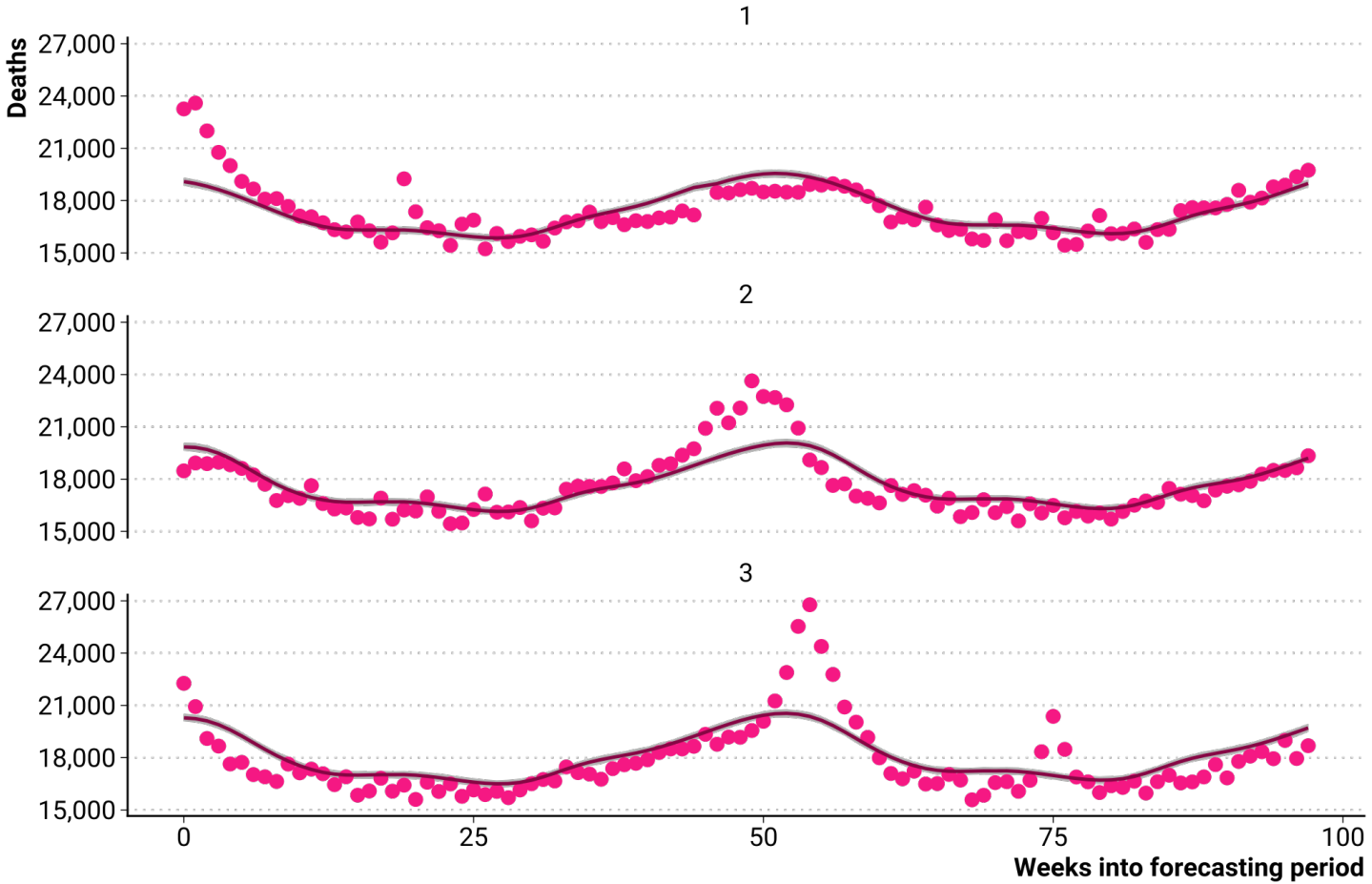
Cross validation – Calibration



Prediction interval with **90% nominal coverage** via Negative-Bin. GAM model

Actual coverage ~87%

Cross validation – Calibration



Prediction interval with **90% nominal coverage** via Poisson GAM model

Actual coverage ~22%

Cross validation – Calibration

Schöley (2022). Empirical prediction intervals. Link to video

Reproducible analysis

github.com/jschoeley/phds22-predictive_modeling

Jonas Schöley

@jschoeley

0000-0002-3340-8518

schoeley@demogr.mpg.de

MAX PLANCK INSTITUTE
FOR DEMOGRAPHIC RESEARCH