Title: Empirical prediction intervals applied to short term mortality forecasts and excess deaths

Authors: Ricarda Duerst[1,2], Jonas Schöley[1]

Affiliations:

[1]Max Planck Institute for Demographic Research, Konrad-Zuse-Straße 1, 18057 Rostock, Germany

[2]University of Helsinki, Yliopistonkatu 3, 00014 Helsinki, Finland

Corresponding authors:

Ricarda Duerst, `duerst@demogr.mpg.de`

Keywords: excess deaths; COVID-19; cross-validation; robustness; empirical prediction intervals

Abstract:

Background: In the Winter of 2022/2023, excess death estimates for Germany indicated a 10% increase, which has led to questions about the significance of this increase in mortality. Given the inherent errors in demographic forecasting, the reliability of estimating a 10% deviation is questionable. This addresses this issue by analyzing the error distribution in forecasts of weekly deaths. By deriving empirical prediction intervals, provide a more accurate probabilistic study of weekly expected and excess deaths compared to the use of conventional parametric intervals.

Methods: Using weekly death data from the Short-term Mortality Database (STMF) for 23 countries, we propose empirical prediction intervals based on the distribution of past out-of-sample forecasting errors for the study of weekly expected and excess deaths. Instead of relying on the suitability of parametric assumptions or the magnitude of errors over the fitting period, empirical prediction intervals reflect the intuitive notion that a forecast is only as precise as similar forecasts in the past turned out to be. We compare the probabilistic calibration of empirical skew-normal prediction intervals with conventional parametric prediction intervals from a negative binomial GAM in an out-of-sample setting. Further, we use the empirical prediction intervals to quantify the probability of detecting 10% excess deaths in a given week, given pre-pandemic mortality trends.

Results: The cross-country analysis shows that the empirical skew-normal prediction intervals are overall better calibrated than the conventional parametric prediction intervals. Further, the choice of prediction interval significantly affects the severity of an excess death estimate. The empirical prediction intervals reveal that the likelihood of exceeding a 10% threshold of excess deaths varies by season. Across the 23 countries studied, finding at least 10% weekly excess deaths in a single week during summer or winter is not very unusual under non-pandemic conditions. These results contrast sharply with those derived using a standard negative binomial GAM.

Conclusion: Our results highlight the importance of well-calibrated prediction intervals that account for the naturally occurring seasonal uncertainty in mortality forecasting. Empirical prediction intervals provide a better performing solution for estimating forecast uncertainty in the analyses of excess deaths compared to conventional parametric intervals.

Introduction

In the Winter 22/23 excess death estimates for Germany were hovering around 10%. Back then, the authors have been contacted by journalists inquiring about the significance of the elevated mortality. Given that statistical estimation always comes with errors attached, can one even reliably estimate a 10% deviation from the norm? In this paper we aim to answer this question via the careful analysis of the error distribution in forecasts of weekly deaths and the seasonality of fluctuations in weekly death counts. Derived from the distribution of errors we propose empirical prediction intervals for the probabilistic study of weekly expected and excess deaths and demonstrate the superior coverage and generality of these intervals compared with conventional parametric intervals. We employ these empirical prediction intervals to quantify, for a range of countries, the probability of observing at least 10% excess death in a given week given the continuation of mortality trends observed prior to the COVID-19 pandemic. Using these p-values we can assess, on a per-country basis, how unusual a 10% mortality increase over the expectation is and whether it should be cause for concern.

Mortality forecasts on a sub-annual timescale have gained relevance as the basis for excess deaths calculations during the COVID-19 pandemic (e.g. [1, 2, 3]). Framed as a forecasting problem, one aims to predict the weekly deaths which would have happened without COVID-19 by forecasting deaths over the pandemic period based on pre-pandemic trends. Those forecast "expected deaths" are associated with an error which can be expressed as a "prediction interval" within which the true expected deaths are to be found with a given probability. It is well known that prediction intervals around forecast values tend to be too narrow (e.g. [4, 5]). In the context of COVID-19 excess death modeling, this phenomenon may lead to wrong conclusions regarding the impact of the pandemic on population mortality, by giving an overly optimistic picture on how precise one can actually forecasts counterfactual weekly expected deaths deaths more than three years post COVID-19. Precisely, given the time passed since the beginning of the pandemic, we expect the uncertainty of forecast expected deaths to have increased. Thus, one might (or should) ask whether we can still reliably detect an, e.g., 10% increase in excess deaths or whether it disappears into the uncertainty of the expected deaths forecasts.

To answer this question, we need reliable measures of forecast uncertainty. It is common practice in demographic forecasting to use parametric prediction intervals, derived from the model structure, such as the random walk over the mortality index of a Lee-Carter model [6], or the Poisson/Negative-Binomial variation around weekly death count [3, 7]. However, of the various sources of error that can contribute to the overall forecast uncertainty (see [8] for an overview), these parametric prediction intervals do not reflect errors from the out-of-sample generalization of the model or inaccurate model specification. This is particularly problematic as the out-of-sample generalization error is the most common type of forecast error [8]. Therefore, the parametric intervals only work if the model is correctly specified and the data generation process does not change over time. As a result, parametric intervals tend to be too narrow.

Empirical prediction intervals, instead on relying on correct parametric specifications, estimate the distribution of error around a forecast value from actual past, out-of-sample forecasting errors. The idea is simple: The error distribution for future forecasts will be similar to the error distribution of past forecasts. Thus, one seeks to estimate the full distribution of forecast error indexed over all desired forecasting strata and time points. In order

to do so, a statistical model is fitted to known out-of-sample forecasting errors. This error distribution can then be used to construct empirical prediction intervals around the central forecast.

Research on empirical uncertainty intervals has been ongoing for at least half a century [9]. The 1980s saw active demographic research on empirical intervals and the authors generally found that the intervals based on past error had better coverage than model-based intervals (e.g. [10, 11, 12]). Since then, research interested in empirical uncertainty has aimed to improve demographic forecasting by quantifying the uncertainty using empirical errors in probabilistic forecasting. A summary of methodological advances (e.g. [8, 13, 14]) and demonstrations of the application of the method can also be found in more recent demographic literature (e.g. [15, 16, 17, 18]).

In the machine learning literature, there is growing interest in "distribution free uncertainty quantification", given that popular techniques do not provide an intrinsic estimate of uncertainty. The community has build a theory on empirical uncertainty intervals under the term "conformal prediction". Here too, the idea is to look at the distribution of historical error measures to assess how likely any given future deviation from the prediction would be [19].

In this paper, we demonstrate the construction and coverage of empirical prediction intervals on the example of excess deaths in 23 countries. The data on weekly deaths for these countries are sourced from the Short-term Mortality Database (STMF, [20]). Further, we compare the calibration of the empirical prediction intervals with a conventional model of parametric prediction intervals. Finally, we answer the application question whether we can reliably detect an 10% increase in excess deaths or whether it disappears into the uncertainty of the expected deaths forecasts, at the current stage of the pandemic.

Data and Methods

Data

We sourced data on weekly deaths from the Short-Term Mortality Fluctuations Database (STMF) [21]. The STMF is part of the Human Mortality Database (HMD) [22] and was established in response to the COVID-19 pandemic and the "increasing importance of short-term or seasonal mortality fluctuations that are driven by temporary hazards such as influenza epidemics, temperature extremes, as well as man-made or natural disasters" [20]. The STMF provides open-access data on mortality by week, sex, and aggregated age group for 38 countries that follow the HMD's criteria of high data quality. The mortality counts are neither adjusted for e.g., under-reporting, nor smoothed [20]. We use weekly death counts for 23 of the available countries. As the data availability across time varies by country, we selected countries whose first observed data entry is for the year 2000 or before.

Further, we use weekly population exposures by country, sex and age in our analysis. These are interpolated from mid-year population estimates from the Human Mortality Database (HMD) [23] using cubic splines and linear extrapolation. For a detailed description of the procedure see [24].

To calculate empirical prediction intervals, perform a validation analysis and an application example, we split the data for each country into three different types of data series (see Figure 1). We use these data series for different parts of our analysis. The first five data series are the calibration series. Using these, we calibrate the empirical prediction intervals. The

following two series are used for the validation of the derived empirical prediction intervals. The forecast period of the last data series spans over the time of the COVID-19 pandemic and is the application series. With this series, we answer our research question on detecting an increase in excess deaths given the uncertainty of the forecasts of expected deaths.

Starting from the last observed week, we split the data into eight partially overlapping series of 365 weeks (7 years). Each series is divided into a training period (blue) of 261 weeks (5 years), and a forecast period (pink) of 104 weeks (2 years) (see Figure 1 for Germany and Table 2 for the starting and end dates of the data series). The series overlap in a way that the forecast period of one series ends right before the forecast period of the following series starts. Therefore, the forecast periods don't overlap each other. This secures that the empirical prediction intervals are not validated on data that is part of several forecast periods. Thus, we avoid the correlation of forecast errors.

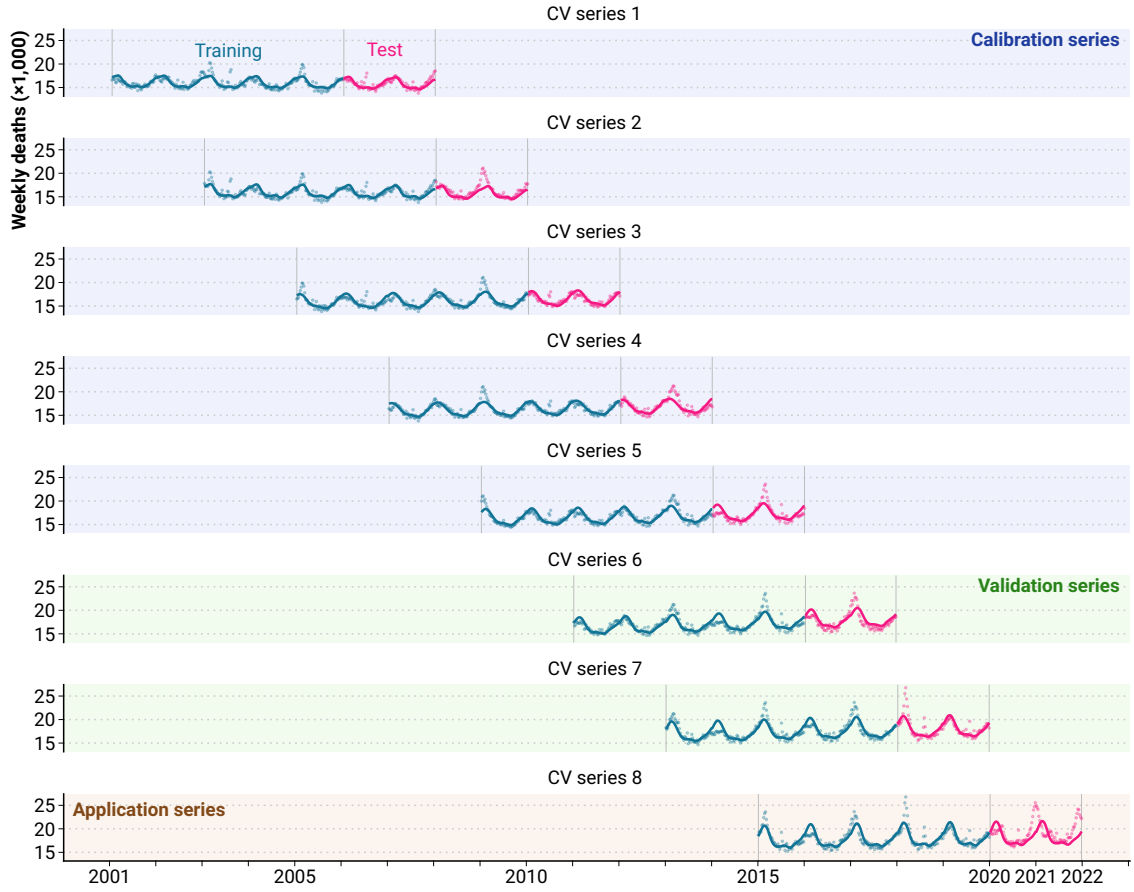

Figure 1: Data-splitting setup. 22 years of weekly death counts have been split into 8 overlapping data series. Each series features 5 years of training data and 2 years of test data. The empirical forecasting error is estimated from the test set of the calibration series, validated on the test set of the validation series and applied to the test set of the application series.

Steps of Analysis

The methodology of our analysis is structured into several steps. With the first four steps, we obtain and calibrate the empirical prediction intervals, and parametric prediction intervals for comparison. For this, we use the calibration data series. During step five, we assess the

calibration of the empirical prediction intervals using the validation data series. Following these steps, we are able to answer our research question regarding the application of empirical prediction intervals to the COVID-19 pandemic.

I Modeling and Forecasting Expected Deaths First, we model and forecast the expected deaths. These are the deaths that would have occurred if a specific event e.g., the COVID-19 pandemic, did not happen. In line with the WHO approach [25], we forecast expected deaths $\hat{y}$ at time $T + h$ using an overdispersed count regression,

$$\hat{y}^E_{T+h} = \exp\left(\alpha + \beta_h h + s_w\left(w\left[h\right]\right)\right), \tag{1}$$

where T is the last time-point in the training data and h is the number of weeks into the forecasting horizon, $\beta_h h$ captures the long term trend of deaths, and $s_w\left(w\left[h\right]\right)$ is a cyclical penalized spline over the week of the year to reflect the seasonality of death counts. The model is fitted to the training periods of the data series and predictions are made over the testing periods. Note that empirical prediction intervals can be derived for any other model for excess deaths, too. For simplicity of exposition we did not include further refinements like population offsets, temperature effects, or auto-correlated errors into the model. We assume the weekly deaths under the expected scenario to be distributed as

$$Y^E_{T+h} \sim \text{Neg.Binomial}(\hat{y}_{T+h}, \phi). \tag{2}$$

II Deriving the Forecast Error In a second step, we quantify the forecasting error at time $T + h$ as the logged ratio between observed and expected deaths over the test period:

$$u_{T+h} = \log \frac{y^O_{T+h}}{\hat{y}^E_{T+h}}. \tag{3}$$

We choose the log-ratio as error scoring function because it reduces the scale dependence and asymmetry of the distribution of errors, in turn making it easier to model the observed error distribution.

III Modeling the Forecast Error Third, we model the time-varying distribution of the observed forecasting error over the forecasting horizon, U_{T+h}, via a skew-normal distribution with a location parameter μ, and time-varying scaling and skewness parameters σ_h and υ_h:

$$U_{T+h} \sim \text{SkewNormal}(\mu, \sigma_h, \upsilon_h). \tag{4}$$

As the observed log-errors are positively skewed, the skew-normal distribution allows for more accurate modeling of the error distribution over time compared to a normal distribution.

We model the scaling parameter as

$$\sigma_h = \exp\left(\alpha_\sigma + \beta_\sigma h + s_\sigma(w[h])\right) \tag{5}$$

where $s_\sigma(w[h])$ is a smooth function of calendar week, capturing the annual seasonality of the error variance. In the context of expected deaths modeling this seasonality is pronounced due to the challenges in predicting the severity of flu-waves during winter. While we would expect the error variance to increase with increasing forecasting horizon as well, we found

that effect to be negligible over the duration of 100 weeks and did not include it here. The skewness υ_h is modeled equivalently via,

$$\upsilon_h = \alpha_\upsilon + s_\upsilon(w[h]), \tag{6}$$

and the mean error is captured in the constant

$$\mu = \beta_{\mu_0}. \tag{7}$$

IV Deriving Prediction Intervals Once estimates for σ_h, μ_h and υ_h are found, in the fourth step, we map the quantiles of the error distribution to the quantiles of the distribution around expected deaths. The p quantile of the distribution of expected/forecasted deaths Y^E_{T+h} can be derived from the corresponding quantile of the error distribution via

$$Q_{Y^E_{T+h}}(p) = \exp(\hat{F}^{-1}_{U_{T+h}}(p)) \cdot \hat{y}^E_{T+h}, \tag{8}$$

where $\hat{F}^{-1}_{U_{T+h}}$ is the estimated quantile distribution of the forecasting error. Intuitively, if we estimated that 95% of the errors in our forecast for a given week should not exceed 1.5 times the central forecast, then the corresponding 95% prediction interval around expected deaths should be 1.5 times the central forecast for that week, likewise for other quantiles.

For comparison of the performance of the empirical prediction intervals with parametric prediction intervals, we derive prediction intervals from a negative-binomial generalized additive model (GAM). The negative-binomial GAM prediction intervals are commonly used for count data (see [26] for examples). Additionally, we calculate the raw quantiles of the forecast error distribution as a second empirical model of the forecast error.

V Assessing Calibration of Empirical Prediction Intervals In the final step, we assess the calibration of the empirical and parametric prediction intervals using two different calibration metrics, the coverage, and the mean interval score.

The coverage measures the fraction of observations that are within the bounds of the prediction interval. Ideally nominal coverage and observed coverage are the same, i.e. a 95% prediction interval should, over the course of many forecasts, cover 95% of realized values. For a collection of prediction intervals (l_i, u_i) and associated observations y_i the coverage can be calculated as

$$\text{Coverage} = \frac{\sum_i^N \mathbf{1}(l_i < y_i < u_i)}{N} \tag{9}$$

where l_i and u_i are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles, $i = 1, \ldots, N$ is an index over all forecasted data points, and $\mathbf{1}(\cdot)$ is the indicator function.

The mean interval score as proposed by Gneiting and Raftery [27] is an example of a *proper scoring rule* and as such has optimal properties for assessing the quality of distributional and, specifically, interval forecasts. In the literature on COVID related prediction models the interval score has seen intensive use [28, 29, 30, 31], and is one of the evaluation measures of the COVID-19 Forecasting Hub Project [32].

The mean interval score balances the two requirements of coverage and sharpness, rewarding narrower prediction intervals but penalizing deviations from the nominal coverage. Among two alternative prediction intervals with equal coverage, the interval which on average is

narrower will yield the lower, i.e. better, mean interval score. This balancing property is very valuable in the evaluation of prediction interval around forecasts of weekly deaths as the variance of the error varies with season. Thus it is possible to have prediction intervals which have perfect coverage over the whole year but are to narrow in winter and too wide in spring. The mean interval score will instead reward prediction intervals which are narrow where they can be, and wide where they must be. The score is defined as follows:

$$S_{\alpha,i}(l_i, u_i, y_i) = \begin{cases} (u_i - l_i) + \frac{2}{\alpha}(l_i - y_i) & \text{for} \quad y_i < l_i \\ (u_i - l_i) + \frac{2}{\alpha}(y_i - u_i) & \text{for} \quad y_i > u_i \end{cases} . \tag{10}$$

We then average the interval scores over all forecasts in the validation series across all countries. Following a suggestion by [33], we calculate the interval scores over log-transformed observations and prediction intervals as otherwise the mean interval score would be dominated by countries with wide intervals (on an absolute expected death scale) due to large population numbers.

Results

Method demonstration on the example of Germany

In the following, we will showcase the application of the empirical prediction interval methodology for Germany. The results of the cross-country analyses will be presented in the following sections.

Following steps I to III of our methodology, Figure 2 shows the distribution of the weekly observed forecast errors in the data series used for calibration, together with modelled forecast error distributions resulting from three different models. Panel (a) shows the observed forecast errors contrasted against the 95% error interval derived from the negative binomial variation, a common parametric specification in excess death modelling [3, 7]. Panels (b) and (c) show two types of post-hoc (empirical) estimates of the forecast error distribution: the skew-normal error model (b) and the raw quantiles of the empirical error distribution (c).

As expected, the observed errors (pink) are positively skewed with a strong seasonal pattern. Both the skewness and the seasonality in the distribution of forecasting errors for Germany, with elevated errors in winter (weeks 0-12, 49-66 and 95-104) and, to a lesser extend, in summer (weeks 20-30 and 75-85), is due to the irregularities in the annual occurrence of flu-epidemics and heat-waves. This seasonal shift in the variance of forecasting error is well reflected by the skew-normal model (b) and the empirical quantiles (c). In contrast, the often used negative binomial specification with fixed overdispersion implies a constant forecasting error (a). Figure 5 in the Appendix shows the forecast error distribution modelled with the skew-normal model for all 23 chosen countries.

As specified in step IV, we derive empirical prediction intervals around the central forecasts of expected deaths from the estimated distribution of forecasting error. For the negative binomial model, we simply report the parametric prediction intervals. Figure 3 contrasts the three approaches in the context of the German COVID-19 pandemic, showing observed and expected weekly deaths since January 2020 along with 95% prediction intervals. The difference between the empirical and parametric prediction intervals is especially notable during winter, where excess deaths seem far more unusual under negative binominal inter-

vals than under the seasonally widened empirical intervals. Notably, and unsurprisingly, the prediction intervals derived from the raw quantiles of the error distribution are not as smooth as the skew-normal modeled errors. Figure 6 shows the empirical skew-normal prediction intervals and observed forecast weekly deaths of the application data series for all 23 countries.

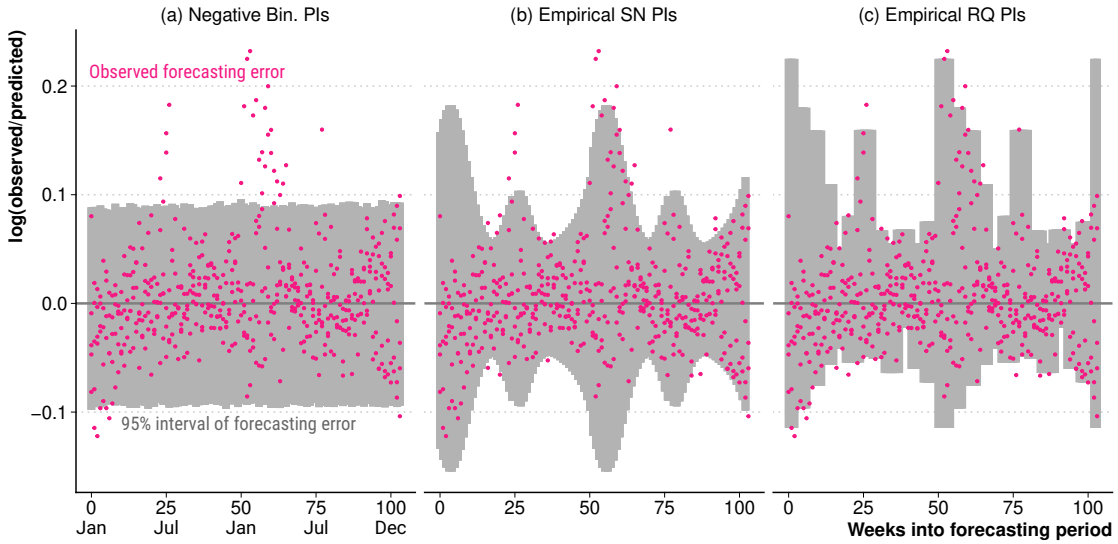

Figure 2: Distribution of the observed forecast errors in the calibration data series (pink), with modelled forecast error distributions, on the example of Germany.

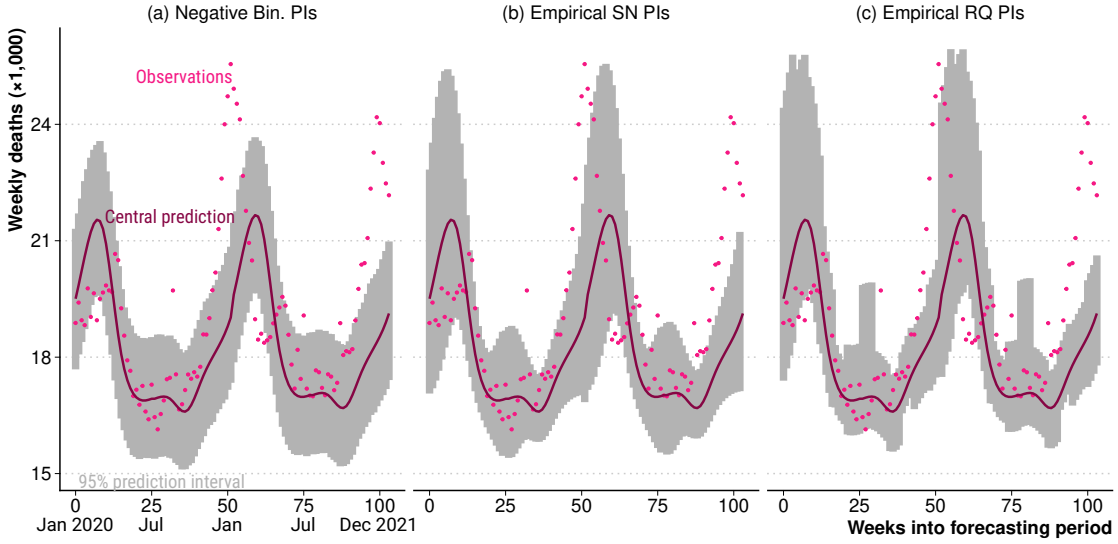

Figure 3: Negative binomial prediction intervals (a) and empirical prediction intervals (b) applied to the application (COVID-19) data series, on the example of Germany.

Cross-country evaluation of forecast calibration

Table 1 shows calibration metrics for the three different prediction interval types as calculated from the validation series. We report the coverage and the mean interval score aggregated across all 23 countries, annually and by season, for a nominal coverage of 95%.

All three types of prediction intervals tend to be too narrow, although close to the nominal coverage. On an annual basis, the negative binomial 95% prediction intervals (PIs) perform best with an actual coverage of 93%, closely followed by the empirical skew-normal PIs with 91%. However, looking at the performance over different seasons, the empirical skew-normal PIs achieve better actual coverage in winter and autumn than the parametric, as the latter is too wide in autumn (99% actual coverage) and too narrow in winter (85% actual coverage). These peculiarities are hidden when only the calibration for the whole year is considered.

These findings are also supported by the mean interval scores. The empirical skew-normal PIs are better calibrated than the negative binomial PIs overall, and in all seasons except summer. This indicates an overall better probabilistic calibration of the empirical skew-normal PIs compared to the conventional negative binomial PIs.

The prediction interval derived from the raw quantiles of the forecast error distribution perform comparatively poorly which may be due to over-fitting. Thus, further results in the paper will be shown without the inclusion of the raw-quantile method.

	Coverage				
	Annual	Dec–Feb	Mar-May	Jun–Aug	Sep–Nov
Negative Bin. PIs	**0.93 (1)**	0.85 (3)	**0.91 (1)**	**0.95 (1)**	0.99 (3)
Empirical SN PIs	0.91 (2)	**0.89 (1)**	0.89 (2)	0.91 (2)	**0.94 (1)**
Empirical RQ PIs	0.86 (3)	0.86 (2)	0.82 (3)	0.86 (3)	0.91 (2)
	Mean Interval Score				
Negative Bin. PIs	0.365 (2)	0.553 (3)	0.376 (2)	**0.287 (1)**	0.248 (2)
Empirical SN PIs	**0.346 (1)**	**0.467 (1)**	**0.369 (1)**	0.310 (2)	**0.241 (1)**
Empirical RQ PIs	0.398 (3)	0.534 (2)	0.437 (3)	0.363 (3)	0.260 (3)

Note: bold font highlights best/same performance among types of prediction intervals
Data Source: Short-term Mortality Fluctuations (STMF) data series, own calculations

Table 1: Calibration metrics for nominal 95% prediction intervals by season and type of interval. The metrics have been calculated on the test periods of the validation data, i.e. on data not seen either during training or calibration, and aggregated over 23 countries.

Probability of 10% excess death under pre-pandemic mortality

The choice of prediction interval is crucial in evaluating the severity of a given excess death estimate. How unusual is a finding of 10% weekly excess deaths in a given country and season given non-pandemic mortality conditions? Figure 4 answers this question for each of the 23 chosen countries under two different prediction intervals, the seasonally adjusted empirical intervals based on past forecast errors and the commonly used negative binomial intervals with time-constant overdispersion. Under the hypothesis that past, non-pandemic mortality trends continue, the graph shows the probability (p-value) for excess deaths in a given week to exceed 10%.

If a standard negative binomial GAM was chosen to derive the prediction intervals, a researcher would hold the belief that exceeding 10% excess death would be just as likely in Spring as in Winter (Figure 4 blue line). The empirical prediction intervals (pink line) shows that this is not the case and that the likelihood of exceeding the 10% threshold varies by

season. A case in point is France, with a peak 20% probability of seeing at least 10% excess deaths for a winter week and a p-value approaching zero during spring. In all countries, there are periods where the p-value exceeds the level of 0.05, yet, there are differences in the seasonal patterns of the p-value plot by country. For most of the countries, the probability of observing 10% excess or more is elevated during the winter and also for the summer weeks (e.g. Austria, Hungary, Portugal, Belgium, Poland, Spain). In contrast, in Norway, Sweden and Finland, we only see a winter and no summer elevation of the p-value. Further, for Croatia and Bulgaria, the p-value to detect 10% excess deaths is even higher in summer than in winter. Therefore, applying to all of the 23 countries, a finding of at least 10% weekly excess deaths in a single week during summer or winter is not very unusual under non-pandemic conditions. This is especially true for smaller populations (e.g. Scotland, Latvia, Luxembourg) where sampling fluctuations in death counts regularly produce excess death estimates above 10% even under normal mortality conditions. In a complementary finding, for Austria, France, Germany, Poland, Spain, and Sweden, the empirical prediction intervals predict a 10% excess p-value close to zero for parts of Spring and Summer, indicating that 10% excess deaths in Spring may well be typical in Winter and cause for alarm in Spring or Fall.

Discussion

Forecasts of weekly deaths have grown in importance since the onset of the COVID-19 pandemic and the need for estimating excess deaths. These forecasts of expected deaths have an error that is probabilistically expressed as a prediction interval. Usually, these prediction intervals are model based parametric derivations. However, they tend to be too narrow, giving a too optimistic outlook on the precision with which one can estimate excess deaths at the current stage of the pandemic. In this paper, we propose the use of empirical prediction intervals for the study of weekly expected and excess deaths and show the superior calibration of these intervals compared to parametric intervals. Further, we demonstrate the application of empirical prediction intervals to the COVID-19 pandemic, answering the question whether a 10% increase in excess deaths can be detected or whether it disappears into the uncertainty of the expected death forecasts.

Empirical prediction intervals are based on modeling the distribution of the forecast errors from past out-of-sample forecast errors. We use a skew-normal distribution with a time varying seasonality parameter to model the forecast errors for weekly deaths, accounting for the positive skewness and seasonality of errors. For the validation analysis, we compare the empirical prediction intervals with parametric prediction intervals obtained from a negative binomial GAM using two measures of calibration: the coverage and the mean interval score. We obtain weekly mortality data from the Short-Term Mortality Database (STMF, [20]) for 23 countries. We showcase the calculation of empirical prediction intervals for these countries, perform the validation analyses, and apply the different intervals to excess deaths in the COVID-19 pandemic.

We showed that our model for the empirical prediction intervals is able to capture the seasonality in weekly deaths and the positive skewness due to the varying interval width over the year. In contrast, the negative binomial prediction intervals have a constant width throughout the year, resulting in excessively wide intervals in Fall and excessively narrow intervals in "inter. Further, our validation analysis demonstrated that the empirical prediction intervals are better calibrated than the negative binomial prediction intervals. This is reflected in the

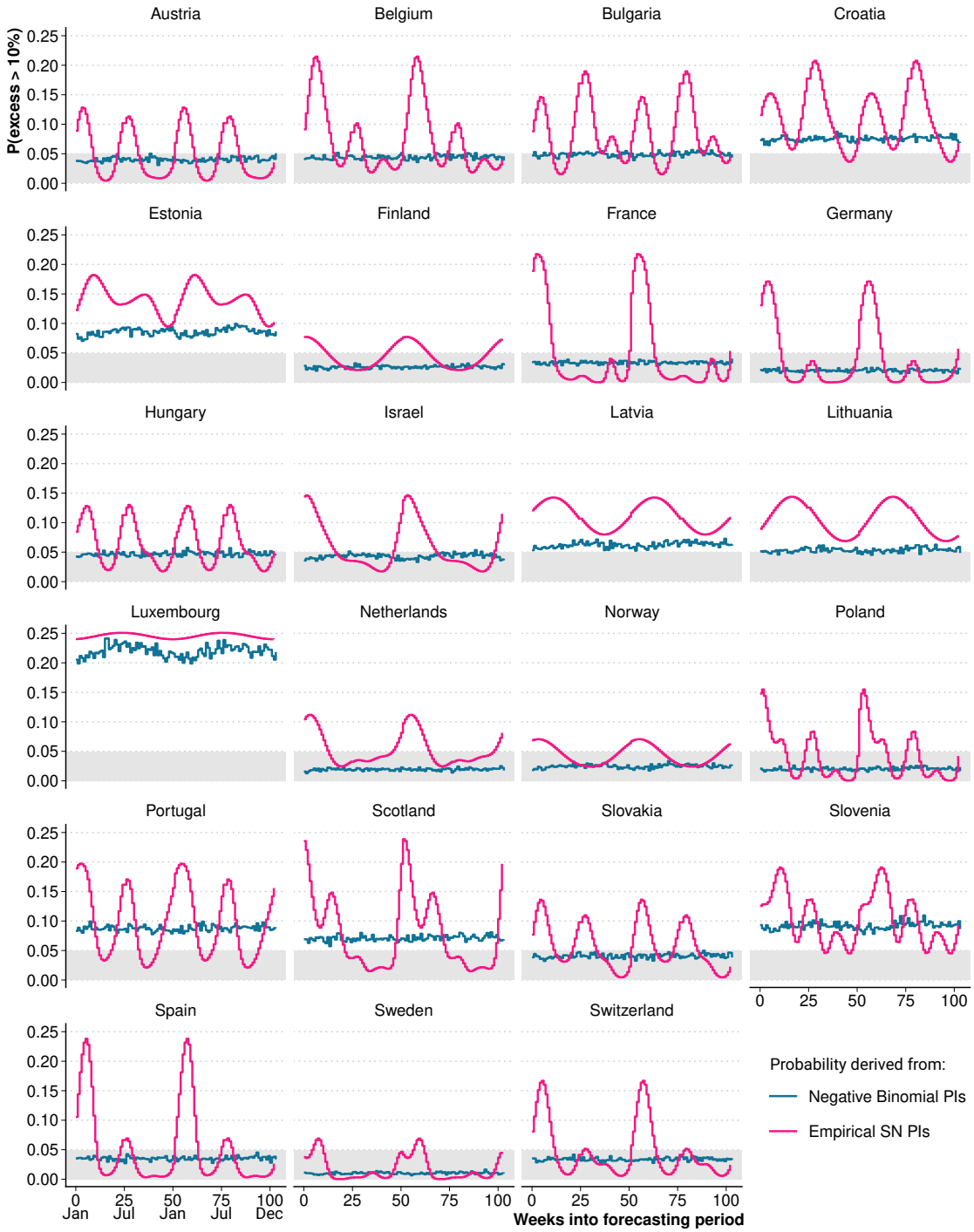

Figure 4: P-value for 10% excess deaths derived from empirical prediction intervals and negative binomial prediction intervals for 23 European countries.

mean interval score that is, on average, better for the empirical prediction intervals. Both types of prediction intervals show a similar good coverage that is, on average, slightly short of the nominal 95%. Regarding the application question, using negative binomial prediction intervals would yield to the result that a 10% increase in excess deaths would be detectable at the current stage of the pandemic. However, the empirical prediction intervals show that this is not the case during the whole year. Regarding the example of Germany, in winter (and with ongoing forecast length also in summer) a 10% in excess deaths due to COVID-19

would disappear into the uncertainty of the forecasts of expected deaths. We found similar results for the other countries in our sample: there are periods throughout the year in which a 10% increase in excess deaths can not reliably predicted for every country. These periods differ between the countries.

The difference in forecast error seasonality between countries may be connected to differences in the prevalent climate. Future research could further investigate this relationship between the climate of a country of interest and the seasonal patterns of uncertainty in mortality forecasting.

Regarding the study of excess deaths due to COVID-19, our results highlight the importance of well calibrated prediction intervals that also address the naturally occurring seasonal uncertainty that comes with mortality forecasting. During winter, the challenge in mortality forecasting lies in judging how severe the flu season is going to be. In summer, the higher uncertainty comes from hard to predict heat waves. These weather-related phenomenons are usually not part of mortality forecast models. Therefore, parametric model-based uncertainty intervals give a too optimistic view of the precision of the expected death forecasts because of their inability to address this seasonal uncertainty. In contrast, the proposed empirical prediction intervals have a superior performance compared to parametric intervals, because they are based on modeling the distribution of the forecast error and thus, their ability to capture the seasonality.

Further, empirical prediction intervals are generally applicable. Not only can they capture different patterns of seasonality in deaths (e.g., high winter uncertainty and low summer uncertainty and vice versa). They can be applied to all models of excess deaths and other methods of forecasting, e.g. forecasts using expert opinions, as they only derive from observable data and not from unobservable model parameters. As long as the forecasts can be validated over a long time, empirical prediction intervals can be used. Therefore, they are not limited to mortality forecasting, either. Using adapted specifications to model the forecast error, they can be applied to e.g., fertility forecasts. Thus, research concerned with demographic forecasting in general can profit from the use of empirical prediction intervals. We invite researchers to use empirical prediction intervals for their forecasting problems and showcase the possibilities of their application.

Data Availability Statement

The research presented in this article is fully reproducible. We have made our codebase available to the scientific community. You can access the complete codebase, including scripts and documentation, at the following link: `https://github.com/jschoeley/epunc`. In addition, the data used for our analyses is sourced from openly accessible and publicly available datasets. Researchers interested in reproducing our work or conducting further investigations can freely download the data from the following link: `https://www.mortality.org/Data/STMF`. The availability of open-access data ensures transparency and promotes collaboration within the scientific community.

Appendix

Table 2: Structure of data used for analyses.

Data Series	Category	Start Date of Training Period	Start Date of Testing Period	End Date of Testing Period
1	Calibration	22.01.2001	23.01.2006	14.01.2008
2	Calibration	20.01.2003	21.01.2008	11.01.2010
3	Calibration	17.01.2005	18.01.2010	09.01.2012
4	Calibration	15.01.2007	16.01.2012	06.01.2014
5	Calibration	12.01.2009	13.01.2014	04.01.2016
6	Validation	10.01.2011	11.01.2016	01.01.2018
7	Validation	07.01.2013	08.01.2018	30.12.2019
8	Application	05.01.2015	06.01.2020	17.12.2021

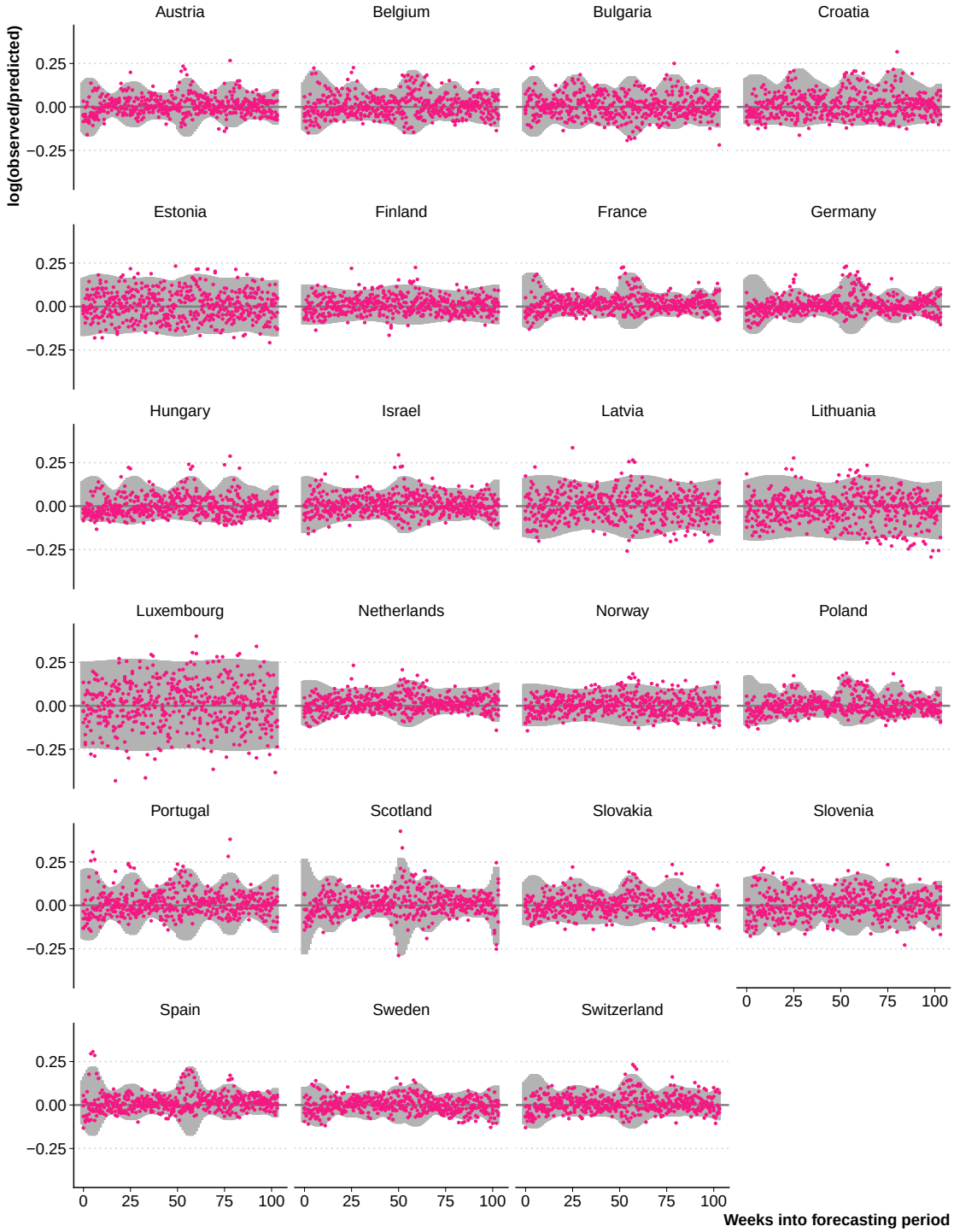

Figure 5: Distribution of the forecast error for weekly death counts as estimated from the calibration data series across 23 European countries.

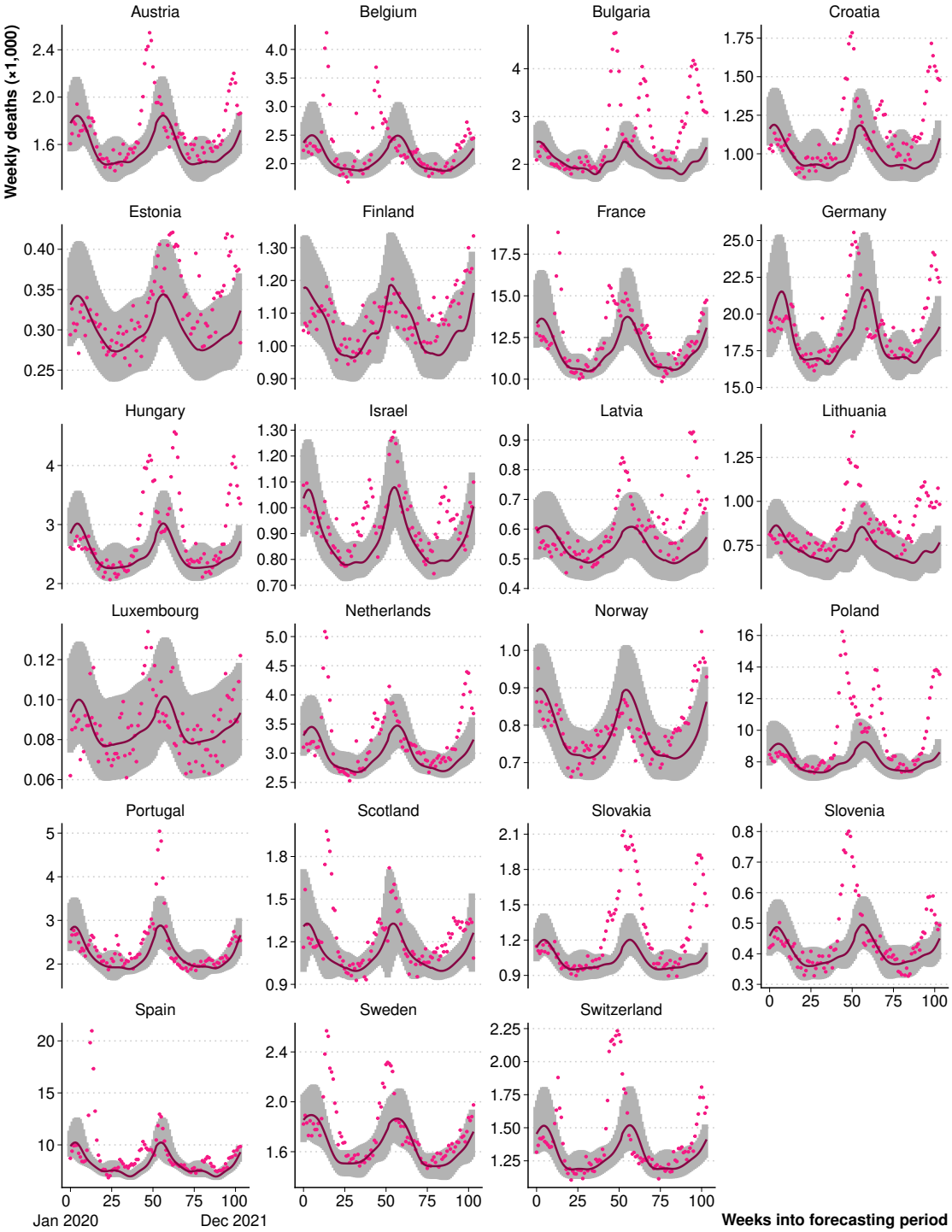

Figure 6: Empirical prediction intervals (95%) applied to the COVID-19 (application) data series across 23 European countries.

References

[1] Vasilis Kontis, James E Bennett, Theo Rashid, Robbie M Parks, Jonathan Pearson-Stuttard, Michel Guillot, Perviz Asaria, Bin Zhou, Marco Battaglini, Gianni Corsetti, et al. Magnitude, demographics and dynamics of the effect of the first wave of the covid-19 pandemic on all-cause mortality in 21 industrialized countries. *Nature medicine*, 26 (12):1919–1928, 2020.

[2] Ariel Karlinsky and Dmitry Kobak. Tracking excess mortality across countries during the covid-19 pandemic with the world mortality dataset. *elife*, 10:e69336, 2021.

[3] Jose Manuel Aburto, Ridhi Kashyap, Jonas Schöley, Colin Angus, John Ermisch, Melinda C Mills, and Jennifer Beam Dowd. Estimating the burden of the covid-19 pandemic on mortality, life expectancy and lifespan inequality in england and wales: a population-level analysis. *J Epidemiol Community Health*, 75(8):735–740, 2021.

[4] Chris Chatfield. Calculating interval forecasts. *Journal of Business & Economic Statistics*, 11(2):121–135, 1993.

[5] Chris Chatfield. Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 158(3):419–444, 1995.

[6] Ronald D Lee and Lawrence R Carter. Modeling and forecasting us mortality. *Journal of the American statistical association*, 87(419):659–671, 1992.

[7] William Msemburi, Ariel Karlinsky, Victoria Knutson, Serge Aleshin-Guendel, Somnath Chatterji, and Jon Wakefield. The who estimates of excess mortality associated with the covid-19 pandemic. *Nature*, 613(7942):130–137, 2023.

[8] Nico W. Keilman. *Uncertainty in national population forecasting: Issues, Backgrounds, Analyses, Recommendations*. Swets & Zeitlinger, Amsterdam, 1990.

[9] W. H. Williams and M. L. Goodman. A simple method for the construction of empirical confidence limits for economic forecasts. *Journal of the American Statistical Association*, 66(336):752–754, December 1971. doi:10.2307/2284223. Publisher: Informa UK Limited.

[10] Michael A. Stoto. The accuracy of population projections. *Journal of the American Statistical Association*, 78(381):13–20, March 1983. doi:10.1080/01621459.1983.10477916. Publisher: Informa UK Limited.

[11] Joel E. Cohen. Population forecasts and confidence intervals for Sweden: a comparison of model-based and empirical approaches. *Demography*, 23(1):105–126, February 1986. doi:10.2307/2061412. Publisher: Duke University Press.

[12] Stanley K. Smith and Terry Sincich. Stability over time in the distribution of population forecast errors. *Demography*, 25(3):461–474, August 1988. doi:10.2307/2061544. Publisher: Duke University Press.

[13] Nico Keilman. Uncertain population forecasts. *Nature*, 412(6846):490–491, 2001.

[14] Yun Shin Lee and Stefan Scholtes. Empirical prediction intervals revisited. *International Journal of Forecasting*, 30(2):217–234, 2014.

[15] Nico Keilman, Dinh Quang Pham, and Arve Hetland. Why population forecasts should be probabilistic-illustrated by the case of norway. *Demographic research*, 6:409–454, 2002.

[16] Nico Keilman and Dinh Quang Pham. Time series based errors and empirical errors in fertility forecasts in the nordic countries. *International Statistical Review*, 72(1):5–18, 2004.

[17] Nico Keilman and Dinh Quang Pham. Empirical errors and predicted errors in fertility, mortality and migration forecasts in the european economic area. *Dicussion Papers*, 386, 2004.

[18] Stefan Rayer, Stanley K. Smith, and Jeff Tayman. Empirical prediction intervals for county population forecasts. *Population Research and Policy Review*, 28(6):773–793, February 2009. doi:10.1007/s11113-009-9128-7. Publisher: Springer Science and Business Media LLC.

[19] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, June 2008. ISSN 1532-4435. doi:10.5555/1390681.1390693. Number of pages: 51 Publisher: JMLR.org tex.issue_date: 6/1/2008.

[20] Dmitri A. Jdanov, Ainhoa Alustiza Galarza, Vladimir M. Shkolnikov, Domantas Jasilionis, László Németh, David A. Leon, Carl Boe, and Magali Barbieri. The short-term mortality fluctuation data series, monitoring mortality shocks across time and space. *Scientific Data*, 8(1), September 2021. doi:10.1038/s41597-021-01019-1. Publisher: Springer Science and Business Media LLC.

[21] László Németh, Dmitri A. Jdanov, and Vladimir M. Shkolnikov. An open-sourced, web-based application to analyze weekly excess mortality based on the Short-term Mortality Fluctuations data series. *PLOS ONE*, 16(2):e0246663, February 2021. doi:10.1371/journal.pone.0246663. Publisher: Public Library of Science (PLoS) tex.owner: jon.

[22] Max Planck Institute for Demographic Research, University of California, Berkeley, and French Institute for Demographic Studies. Human mortality database, 2023. URL `https://mortality.org/`.

[23] Magali Barbieri, John R Wilmoth, Vladimir M Shkolnikov, Dana Glei, Domantas Jasilionis, Dmitri Jdanov, Carl Boe, Timothy Riffe, Pavel Grigoriev, and Celeste Winant. Data resource profile: The human mortality database (HMD). *International Journal of Epidemiology*, 44(5):1549–1556, June 2015. ISSN 0300-5771. doi:10.1093/ije/dyv105. URL `https://doi.org/10.1093/ije/dyv105`. Publisher: Oxford University Press (OUP).

[24] Jonas Schöley. Robustness and bias of European excess death estimates in 2020 under varying model specifications. *medRxiv*, June 2021. doi:10.1101/2021.06.04.21258353. Publisher: Cold Spring Harbor Laboratory.

[25] Daniel M Weinberger, Jenny Chen, Ted Cohen, Forrest W Crawford, Farzad Mostashari, Don Olson, Virginia E Pitzer, Nicholas G Reich, Marcus Russi, Lone Simonsen, et al.

Estimation of excess deaths associated with the covid-19 pandemic in the united states, march to may 2020. *JAMA internal medicine*, 180(10):1336–1344, 2020.

[26] David J Olive, Rasanji C Rathnayake, and Mulubrhan G Haile. Prediction intervals for glms, gams, and some survival regression models. *Communications in Statistics-Theory and Methods*, 51(22):8012–8026, 2021.

[27] Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. doi:10.1198/016214506000001437. Publisher: Informa UK Limited.

[28] Logan C Brooks, Evan L Ray, Jacob Bien, Johannes Bracher, Aaron Rumack, Ryan J Tibshirani, and Nicholas G Reich. Comparing ensemble approaches for short-term probabilistic covid-19 forecasts in the us. *International Institute of Forecasters*, 2020.

[29] Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2):e1008618, 2021.

[30] Johannes Bracher, Daniel Wolffram, Jannik Deuschel, Konstantin Görgen, Jakob L Ketterer, Alexander Ullrich, Sam Abbott, Maria Vittoria Barbarossa, Dimitris Bertsimas, Sangeeta Bhatia, et al. A pre-registered short-term forecasting study of covid-19 in germany and poland during the second wave. *Nature communications*, 12(1):5173, 2021.

[31] Katharine Sherratt, Hugo Gruson, Helen Johnson, Rene Niehus, Bastian Prasse, Frank Sandmann, Jannik Deuschel, Daniel Wolffram, Sam Abbott, Alexander Ullrich, et al. Predictive performance of multi-model ensemble forecasts of covid-19 across european nations. *Elife*, 12:e81916, 2023.

[32] Estee Y Cramer, Yuxin Huang, Yijin Wang, Evan L Ray, Matthew Cornell, Johannes Bracher, Andrea Brennen, Alvaro J Castro Rivadeneira, Aaron Gerding, Katie House, Dasuni Jayawardena, Abdul H Kanji, Ayush Khandelwal, Khoa Le, Jarad Niemi, Ariane Stark, Apurv Shah, Nutcha Wattanachit, Martha W Zorn, Nicholas G Reich, and US COVID-19 Forecast Hub Consortium. The united states covid-19 forecast hub dataset. *medRxiv*, 2021. doi:10.1101/2021.11.04.21265886. URL `https://www.medrxiv.org/content/10.1101/2021.11.04.21265886v1`.

[33] Nikos I Bosse, Sam Abbott, Anne Cori, Edwin van Leeuwen, Johannes Bracher, and Sebastian Funk. Scoring epidemiological forecasts on transformed scales. *PLoS Computational Biology*, 19(8):e1011393, 2023.