Robustness and bias of European excess death estimates in 2020 under varying model specifications

Jonas Schöley[1*],

1 Laboratory of Population Health, Max Planck Institute for Demographic Research, Rostock, Germany

* schoeley@demogr.mpg.de

Abstract

Various procedures are in use to calculate excess deaths during the ongoing COVID-19 pandemic. Using weekly death counts from 20 European countries, I evaluate the robustness of excess death estimates to the choice of model for expected deaths and perform a cross-validation analysis to assess the error and bias in each model's predicted death counts. I find that very different model specifications produce similar patterns of weekly excess deaths but yield substantially varying levels of excess. While the exact country ranking along percent excess death in 2020 is sensitive to the choice of model the top and bottom ranks are robustly identified. On the country level, the 5-year average death rate model tends to produce the lowest excess death estimates, whereas high excess deaths are produced by the popular 5-year average death count and Euromomo-style Serfling models. Cross-validation revealed these estimates to be biased under a causal interpretation of "expected deaths had COVID-19 not happened". Sophisticated regression models controlling for autocorrelation and temperature effects did not improve prediction error over simpler specifications.

Introduction

The concept of "excess deaths" became widely known in 2020. Early in the year, international newspapers reported weekly death counts that were notably elevated compared to previous years [1,2]. In a situation where tests for COVID-19 were scarce and the underlying cause of death often hard to determine, excess deaths proved to be a rational means of monitoring the direct and indirect mortality impact of the ongoing COVID-19 pandemic with minimal data requirements [3]. Since the start of the pandemic, the literature on excess deaths has been growing fast and now features international comparisons and rankings owing to the readily available data on death counts for many countries [4–9]. However, despite the prominence and relevance of the measure, the excess death methodology is far from standardized with different definitions of excess deaths and different models used to estimate an expected death counts against which the excess is judged. This eclectic state of the art begs the questions on robustness and bias of excess death estimates and cross-country comparisons under various analysis strategies.

Given data from 20 European countries, I assess how excess death measures in the year 2020 vary under different modeling choices for the expected number of deaths. Furthermore, the models are tested for bias in their predicted weekly and annual death counts using a time-series cross-validation setup. While the lack of consensus in methodology extends beyond the model specification to the very definition of excess deaths, I focus entirely on the impact of model specification and define excess as observed minus expected, allowing the measure to be negative.

As the literature on excess deaths in 2020 is dominated by simple average or regression-based approaches, the scope of this paper is restricted to different regression specifications of increasing elaboration. Models based upon multi-year averages are prominently employed by newspapers and statistical offices and featured in the academic literature [10–13]. While simple and easy to communicate, these procedures do not adjust for time trends in the death counts or rates. This issue is addressed by the Serfling model [14], which in various specifications has been applied to quantify COVID-19 related deaths [9,15–18]. Generalized additive models relax the strict specifications of the Serfling model and allow for smooth long-term and seasonal

effects [19,20]. These models may be further elaborated via the inclusion of temperature effects, autoregressive residuals, adjustments for bank holidays, and Bayesian inference [4].

Having chosen 9 model specifications reflecting the heterogeneity in the literature, each model is then probed for a tendency to produce high or low estimate of expected and conversely excess deaths compared with the ubiquitous average weekly death count model. I check if the ranking of European countries along the total percentage of excess deaths in 2020 differs under different baseline models, and identify those countries where there is disagreement about the pattern or existence of excess deaths among the models.

When excess death numbers vary by model, the question arises which model to trust. In line with [4] and [21] I argue that the challenge of estimating the expected number of weekly deaths in the year 2020, given that COVID had not occurred, is a classical forecasting challenge, and thus, any model can be validated by testing its predictions on past data. To that end, I construct a cross-validation challenge, where each model is fitted on seven years of weekly death counts and is then tasked to predict the next 52 weeks from that series. The error and bias of any model can then be estimated by comparing the model predictions with the observed data. If a model's predictions over past years tend to over/underestimate observed death counts, then this model can be expected to also over/underestimate the death count in 2020 under the counterfactual no-COVID scenario, thereby biasing excess deaths.

Materials and Methods

Raw and derived data

To replicate the more elaborate expected death regression models, I collected cross-country data on weekly death counts and population exposure by age and sex, along with information on the timing of public holidays and weekly population-weighted temperature anomalies (S1 Fig). Computer code to replicate this data collection and the processing steps described below is available at `github.com/jschoeley/mocy`.

Age and sex-specific weekly death counts by country were sourced from the Short-term Mortality Fluctuations (STMF) database [22,23]. European countries with

data going back to at least 2007 were included in the analysis. As the Mean Absolute ₆₁
Percentage Error (MAPE) measure is used to report on prediction error, strict ₆₂
positivity in the weekly death counts was required. I thus excluded Estonia, Iceland, ₆₃
and Luxembourg from the analysis and summed deaths into age categories 0 to 65, 65 ₆₄
to 75, 75 to 85, and 85+. ₆₅

Person-weeks of population exposure by country, sex, and age were derived from ₆₆
mid-year population estimates from the Human Mortality Database [24]. I summed ₆₇
these estimates into the target age groups and then interpolated over the weeks of a ₆₈
year by fitting a cubic spline and extrapolating linearly to the end of 2020. Integrating ₆₉
the spline over single weeks yielded person-weeks of exposure. ₇₀

Population weighted weekly temperature anomalies were calculated from global ₇₁
gridded temperature [25] and population data [26]. First, I calculated the average ₇₂
weekly temperature per grid cell from the daily measurements and then weighted these ₇₃
measurements by the population size in the same grid cell. I then selected all grid-cells ₇₄
overlapping with the area of a given country and calculated the country's ₇₅
population-weighted average weekly temperature. Averaging the weekly temperatures ₇₆
over multiple years gave an expected temperature. Country-specific weekly ₇₇
population-weighted temperature anomalies were then derived by subtracting this ₇₈
temperature expectation from the observed weekly temperature in a given year. ₇₉

The `Nager.Date` software [27] was employed to determine for each week, year, and ₈₀
country the occurrence and type of a public holiday, distinguishing between Christmas, ₈₁
Easter, New Year, and "other" types of public holidays. ₈₂

Model choice ₈₃

Expected deaths and associated excess deaths statistics were computed under nine ₈₄
different models listed in Table 1. The models were chosen to represent the range of ₈₅
approaches in the literature on excess deaths estimation since the start of the pandemic, ₈₆
from very simple to elaborate. All models predict age and sex specific expected death ₈₇
counts and are fitted separately for each of the 20 countries. ₈₈

With the exception of the Latent Gaussian Models (LGMr, LGMrt) all models were ₈₉
specified as quasi-Poisson regressions on weekly death counts D_i of the form ₉₀

Table 1. Models for weekly expected deaths.

Model	Description
Weekly averages (AVGc5) 5 year average death counts (AVGr5) 5 year average death rates	Average over the weekly deaths counts or rates in the preceding 5 years implemented as quasi-Poisson GLM with week-of-year coefficients
Serfling Model (SRFc) without exposures (SRFcem) Euromomo style, i.e. no exposures, fitted only on weeks without flu-activity (SRFr) with exposures	Quasi-Poisson regression on death counts with log-linear long term trend, AIC selected Fourier-term seasonality, and public-holiday coefficients
Generalized Additive Model (GAMr) without temperature anomaly predictor (GAMrt) with temperature anomaly predictor as smoothly varying coefficient over week of year	Quasi-Poisson regression on death counts with log-linear long term trend, penalized cyclical spline seasonality, and public-holiday coefficients
Latent Gaussian Model (LGMr) without temperature anomaly predictor (LGMrt) with temperature anomaly predictor as varying coefficient over week of year implemented as cyclic random walk	Bayesian Poisson regression on death counts with autoregressive trend, non-parametric time-varying seasonality, and public-holiday coefficients

$$D_i \sim \text{Pois}(\lambda_i, \phi)$$
$$\lambda_i = \exp(\eta_i), \tag{1}$$

where λ_i is the expected death count, ϕ is the dispersion factor, and η_i is the linear predictor on a log-scale. While we considered employing the negative binomial distribution instead, convergence issues due to sporadic underdispersion informed our choice of the more robust quasi-Poisson model.

For the *5 year average death count model* (AVGc5) we specified

$$\eta_i = \alpha_{j[i]w[i]}, \tag{2}$$

where $\alpha_{j[i]w[i]}$ is a coefficient for iso-week w and sex×age stratum j of observation i within some country. The model was fit over a period of five years.

Log-person-weeks exposure $\log(E_i)$ were added as an offset to Eq 2, yielding the *5*

year average death rate model

$$\eta_i = \alpha_{j[i]w[i]} + \log(E_i). \tag{3}$$

The *Serfling with exposures* (SRFr) model was specified as

$$
\begin{aligned}
\eta_i =& \alpha_{j[i]} + \beta_{j[i]}t_i + \\
& \gamma_{1,j[i]}\sin\left(\frac{2\pi}{52}w_i\right) + \gamma_{2,j[i]}\cos\left(\frac{2\pi}{52}w_i\right) + \gamma_{3,j[i]}\sin\left(\frac{2\pi}{26}w_i\right) + \gamma_{4,j[i]}\sin\left(\frac{2\pi}{26}w_i\right) + \\
& \delta_{j[i]k[i]} + \log(E_i).
\end{aligned}
\tag{4}
$$

Eq 4 features log-linear time trends $\beta_{j[i]}$ over weeks t_i since the earliest observation, parametric seasonal effects $\gamma_{j[i]}$, and the effect of a public holiday $\delta_{j[i]k[i]}$ on the expected death counts, where k denotes the type of holiday. The model is completely stratified over sex and age groups j. Within each stratum the number of seasonal terms (no seasonality or $\gamma_{1,2}$ or $\gamma_{1,2,3,4}$) were selected via AIC using the likelihood from an initial Poisson regression without overdispersion. Omitting the term $\log(E_i)$ from Eq 4 yields the *Serfling without exposures* (SRFc) model. For the *Euromomo style Serfling model* (SRFcem) the exposure was dropped as well and the fitting period truncated to weeks 15 through 26 and 36 through 45. All Serfling models were fitted via the `glm` function in `R`.

The *Generalized Additive Model with temperature anomaly* (GAMrt) allows for non-parametric, smooth seasonal effects and adjusts for the effect of extreme temperature on weekly death counts. I specified the model as

$$\eta_i = \alpha_{j[i]} + \beta_{j[i]}t_i + \gamma_{j[i]w[i]} + \nu_{w[i]}T_i + \delta_{j[i]k[i]} + \log E_i, \tag{5}$$

where $\gamma_{j[i]w[i]} = f_{j[i]}(w_i)$ is a stratum specific seasonality term implemented as a cyclical penalized spline with 12 knots spaced in equal distance over the weeks of a year w_i, and $\nu_{w[i]} = g(w_i)$ is the temperature anomaly coefficient varying smoothly over w_i in a cyclical fashion. For the *Generalized Additive Model without temperature anomaly* (GAMr) this term was omitted. Analogous to the Eq 4 I specified a log-linear trend, public holiday intercepts, and an offset controlling for population exposure. The model

was fitted via the `gam` function from the `mgcv` library in `R`. ⟨120⟩

The *Latent Gaussian Model with temperature anomaly* (LGMrt) follows the ⟨121⟩ specification by [4]. It features a nonparametric seasonal effect which may vary over ⟨122⟩ time and an autoregressive trend. The model is fitted separately by age group and sex ⟨123⟩ and, omitting the subscript $j[i]$, can be written as ⟨124⟩

$$D_i \sim \text{Pois}(\lambda_i)$$
$$\lambda_i = \exp(\eta_i) \qquad\qquad (6)$$
$$\eta_i = \alpha + \beta t_i + \zeta_{t[i]} + \gamma_{t[i]} + \nu_{w[i]} T_i + \delta_{k[i]} + \varepsilon_i + \log E_i,$$

where βt_i is a simple log-linear trend over time and $\zeta_{t[i]} \sim \text{Norm}(\varphi \zeta_{t[i]-1}, \sigma_\zeta^2)$ is a ⟨125⟩ first order autoregressive trend. In contrast with the other models over-dispersion is ⟨126⟩ directly included in the linear predictor in form of the unstructured error term ⟨127⟩ $\varepsilon \sim \text{Norm}(0, \sigma_\varepsilon^2)$. Seasonality is captured with a time-varying random coefficient $\gamma_{t[i]}$, ⟨128⟩ following [28] section 3.5, where the summed effect of every consecutive series of 52 ⟨129⟩ weeks on mortality is distributed as $\sum_{s=0}^{51} \gamma_{t[i]+s} \sim \text{Norm}(0, \sigma_\gamma^2)$. The effect of ⟨130⟩ temperature anomalies on weekly mortality is captured by the varying coefficient ⟨131⟩ $\nu_{w[i]} \sim \text{Norm}(\nu_{w[i]-1}, \sigma_\nu^2)$, specified as a circular first order random walk over week of ⟨132⟩ year. The temperature effect was dropped for the *Latent Gaussian Model without* ⟨133⟩ *temperature anomaly* (LGMr). The models were fitted with the `INLA` library [29, 30] in `R`. ⟨134⟩

Excess death measures ⟨135⟩

Excess deaths D^* in a given week t and stratum j are defined as $D_{tj}^* = D_{tj} - \lambda_{tj}$, ⟨136⟩ where D is the observed death count and λ are the expected deaths under a given ⟨137⟩ model. This definition as a model residual allows to capture mortality displacement ⟨138⟩ effects where a positive excess in part of the year can be followed by negative excess in ⟨139⟩ later weeks. The total number of excess deaths over period $[a, b]$ are given by ⟨140⟩ $D_{[a,b],j}^* = D_{[a,b],j} - \lambda_{[a,b],j}$, where $D_{[a,b],j} = \sum_{t=a}^{b} D_{tj}$ and $\lambda_{[a,b],j} = \sum_{t=a}^{b} \lambda_{tj}$. Following ⟨141⟩ a common strategy, expected and observed deaths were summed over strata to estimate ⟨142⟩ the excess at a higher aggregation level [4, 15, 16, 19]. ⟨143⟩

To facilitate comparisons between countries, I report the weekly percentage increase ⟨144⟩ of observed deaths over expected deaths, $P_{tj} = \frac{D_{tj}^*}{\lambda_{tj}}$, the so-called "P-score". For a ⟨145⟩

sequence of weeks $[a, b]$ a "cumulative P-score" was formed, calculated from the total excess and expected deaths over that period, $P_{[a,b],j} = \frac{D^*_{[a,b],j}}{\lambda_{[a,b],j}}$.

Prediction intervals

Prediction intervals around the measures of excess death were computed via simulation. First, 500 sets of parameter estimates, $s = \{1, \ldots, 500\}$, were sampled either from a multivariate normal distribution $\text{Norm}(\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\boldsymbol{\beta}$ being the vector of estimated parameters, and $\boldsymbol{\Sigma}$ the variance-covariance matrix of the fitted model, or, in case of the Latent Gaussian Models, from the posterior distribution of the parameters. For each of the 500 simulated sets of parameters I calculated the expected weekly death count λ^s_{tj} and then sampled a single simulated death count d^s_{tj} from a negative-binomial distribution with mean equal to λ^s_{tj}, and variance equal to $\lambda^s_{tj}\phi$, matching the characteristic moments of the quasi-Poisson distribution. In case of under-dispersion, where $\phi < 1$, and for the Latent Gaussian Models, I sampled from a Poisson distribution with rate λ^s_{tj}. The simulated death counts d^s_{tj} were then substituted into the excess death formulas in place of the expected deaths λ_{tj} and the resulting distribution of excess summarized by its quantiles.

Cross-validation

In order to assess the accuracy of the predicted death counts from each model a cross-validation analysis was performed. Time series of weekly death counts and co-variates were split into five cross-validation sets (S2 Fig). The training period starts at week 27 of a year and ends after seven fully observed summer to summer seasonal cycles with the beginning of week 8 of a year. After fitting the model, weekly death counts were predicted for 45 weeks following week 8 of the last year in the training set. The starting years for the five cross-validation series were 2007 to 2011. This particular cross-validation setup has been chosen to mirror the task of predicting weekly deaths in 2020, given data observed until shortly before the outbreak of the pandemic.

I compared the expected death counts λ_i with the observed death counts D_i in the test set and quantified the prediction bias via the mean percentage error, $\text{MPE} = 1/n \sum_{i=1}^{n} \frac{D_i - \lambda_i}{D_i}$, and the prediction error via the mean absolute percentage

error, MAPE $= 1/n \sum_{i=1}^{n} \frac{|D_i - \lambda_i|}{D_i}$, where i runs over the n predictions in the five test   175
sets.   176

All predictions were made by country, week, sex, and age. To facilitate presentation   177
I first summed the observed and predicted death counts a) over weeks, sex, and age   178
(annual deaths by country), and b) over sex and age alone (weekly deaths by country),   179
then calculated the aggregated prediction bias and error by country, and finally   180
summarized the distribution of these aggregated measures of model fit across countries   181
via quantiles. Stratum-specific prediction errors can be found in the appendix.   182

Results   183

Fig 1. Weekly and cumulative excess death percentage. Shown are estimates
from 9 different models during the year 2020 weeks 8 through 52 for 20 European
regions. Points indicate weekly excess percentages. Cumulative excess percentages are
indicated by curves and derived from cumulated observed and expected deaths. The
labels refer to the range of predicted percent excess over the entire analysis period.
Y-axis truncated at 60%.

Models agreed on the weekly *pattern* of excess deaths but differed in the estimated   184
level of excess (Fig 1). The following weekly *patterns* were identified by all models: 185
Austria, Switzerland and the Eastern European countries suffered higher weekly peak   186
excess deaths in fall and winter than in spring. The situation was reversed in France,   187
Netherlands, Scotland, Spain, and Sweden, with lower levels of weekly percent excess in   188
the second half of the year. Belgium and Portugal experienced three distinct waves of   189
pronounced excess death in 2020. Disagreement among models regarding the *level* of   190
country specific weekly and annual percent excess death in 2020 was typically around   191
5%, but could reach double digits within some age and sex strata (S3 Fig–S10 Fig).   192
Consequently, models disagreed at times regarding the existence of significantly elevated   193
annual percent excess ($P_{2020} > 0$ with probability ≥ 0.95). On the country level such   194
disagreement only manifested for Denmark, Finland, Latvia, and Norway, countries   195
with relatively few registered COVID deaths. Within age and sex strata model   196
disagreement was more widespread, affecting 15 out of 20 countries. Belgium, Spain,   197
Scotland, Portugal, and Slovenia were the only countries where all nine models agreed,   198
within *all* age and sex strata, that annual percent excess death has been significantly   199

elevated above zero in 2020. On the country level the average death rate model (AVGr5) 200
produced the lowest annual percent excess death estimates in 18 out of 20 countries, 201
whereas the maximum varied, with the Euromomo style Serfling (SRFcem) and the 202
average death count model (AVGc5) producing the highest excess in 13 out of 20 203
countries. 204

Fig 2. Country ranking of excess death percentage during the year 2020 weeks 8 through 52 under 9 different models.

The ranking of countries along the percent excess death in 2020 changed between 205
models (Fig 2). All models agreed on the five countries with the highest annual percent 206
excess (Poland, Spain, Belgium, Slovenia, Czech Republic) and on the four least affected 207
countries (Finland, Latvia, Denmark, Norway). There was substantial disagreement 208
regarding the middle ranks with Hungary, Lithuania, Sweden, Slovakia, Austria, 209
Hungary, Netherlands, and Portugal jumping by more than 5 positions between models. 210
Ranks under the Serfling with exposures (SRFr) and Generalized Additive Models 211
(GAMr, GAMrt) exhibited high agreement. Within age and sex strata more drastic 212
jumps in rank across models could be observed, but again with good agreement 213
concerning the top and bottom ranks (S11 Fig). 214

Fig 3. Percent differences of predicted death counts from various models against the 5-year average weekly death count model prediction. The prediction differences were summarized across countries by the 0.25, 0.5, and 0.75 quantiles.

I compared the 5-year average death count model (AVGc5, the reference) to various 215
other models concerning the expected number of deaths in 2020 (Fig 3). With the 216
exception of the Euromomo Serfling model (SRFcem), all models tended to predict 217
higher numbers of weekly and annual deaths compared to the average death count 218
(AVGc5) model. The 5-year average death rate model (AVGr5) gravitated towards the 219
highest expected death counts, with annual predictions being higher by 3.1 to 5.1% 220
compared to the reference for half of the countries. Rather high baseline estimates were 221
found for the Latent Gaussian Models (LGMr, LGMrt) and the Serfling without 222
exposures (SRFc) model. Within population strata results varied. The tendency for the 223
AVGc5 model to produce low expected death counts was most pronounced in the age 224
group 85+, but clearly reversed in the age group 0 to 65. Likewise, the tendency for the 225

AVGr5 model to produce high estimates of excess was present in all age groups but the   226

highest (S12 Fig).   227

Fig 4. Bias and error by model when predicting death counts on test data.
The error (MAPE, italic) and bias (MPE, bold) measures were summarized across
countries by the 0.25, 0.5, and 0.75 quantiles.

The cross-validation study confirmed the substantial biases of the Serfling   228

Euromomo regression and the 5-year average death rate and count models (Fig 4).   229

While the average rate model (AVGr5) consistently displayed the highest propensity   230

among all models to overestimate annual or weekly deaths on the country level, the   231

average count (AVGc5) and the Serfling Euromomo models (SRFcem) underestimated   232

deaths. The lowest bias was presented by the models GAMr and SRFr. For both models   233

the average percentage error when predicting annual death counts was between -1% 234

and 0.2% for half of the countries with similar results for week-by-week predictions.   235

Omitting the population structure term from a Serfling model increased the prediction   236

bias as seen for model SRFc. Between sex and age strata the relative model   237

performance when predicting annual deaths often varied with model AVGc5 changing   238

direction of the bias over age and model SRFcem being comparably unbiased in the age   239

group 85+ (S13 Fig). Models GAMr and SRFr consistently displayed low bias across   240

strata. I found no evidence that the inclusion of a temperature variable decreased the   241

prediction bias. The Latent Gaussian models (LGMr, LGMrt) exhibited a tendency for   242

over-prediction, but less severe than seen with model AVGr5.   243

Prediction errors tended to more than double going from annual to weekly   244

predictions and varied across countries (Fig 4). Annual death counts were predicted   245

with a MAPE ranging from 1.9–3.4% across models, increasing to 4.2–5.9% for the   246

weekly predictions. Notably, the annual country level predictions from the simple AVGc   247

and the biased SRFcem model exhibited low MAPEs. This result however, did not   248

extend to annual predictions by age and sex, where both models showed varying biases   249

and comparatively high errors (S13 Fig). Because the annual predictions were   250

aggregated from the stratum specific predictions, this suggests a potential for errors to   251

cancel out in the aggregation process. Within strata the GAM and SRFr models   252

exhibited the lowest error when predicting annual deaths and the relative disadvantage   253

of the LGM models increased with age. Omitting the exposure variable from the   254

Serfling model increased the MAPE. The inclusion of a temperature effect did not   255
consistently decrease the prediction error.   256

Discussion   257

No matter the chosen model, in 2020, deaths across Europe have been unusually   258
elevated, exhibiting a weekly pattern closely following the course of the COVID-19   259
pandemic in the different countries. Other inferences, however, are less robust to the   260
choice of model. Depending on how the expected deaths were calculated, one may have   261
either Poland or Spain as the worst affected country, Denmark may or may not have   262
had significantly elevated death counts, annual deaths among Slovakian males ages 85+   263
may be increased by as little as 4% to as much as 15%, and death counts following the   264
first wave of the pandemic may have returned to expected levels or dropped into a   265
"mortality deficit". Thus, short of a standard procedure of specifying expected deaths,   266
robustness checks are crucial. In the simplest case, such a check could be based upon   267
expected deaths via 5-year average weekly death rates and alternatively a Serfling   268
model similar to the Euromomo implementation, as we found these two approaches to   269
have opposite biases, with predictions from other methods often falling in-between.   270

   The Euromomo model is designed to predict death counts as expected under   271
exceptionally mild seasonal mortality increases, thus resulting in high estimates of   272
excess death. While this makes the Euromomo approach very useful for monitoring   273
influenza-related mortality, the baseline is biased if interpreted as "expected deaths in   274
the absence of the COVID-19 pandemic", as the typical seasonal increase in deaths is   275
not as shallow as predicted under Euromomo. The technical reason for the low   276
Euromomo predictions during fall and winter is the exclusion of weeks with likely   277
influenza activity from the model fitting. If the modeling goal is mortality forecasting,   278
as under the counterfactual no-COVID scenario, then fitting seasonal regression models   279
over all weeks of a year reduces the bias in the forecasted death counts.   280

   A lack of adjustment for continuous mortality improvements, as have been observed   281
across Europe up until 2020 [31,32], explains the tendency of the average death rate   282
model towards low estimates of excess. By simply averaging declining death rates over   283
time, one overestimates expected mortality for the current year, thus underestimating   284

excess. Claims of no excess deaths based upon averaged weekly death rates must therefore be checked for robustness against the inclusion of a time trend.

Two opposing time-trends explain the bias of the ubiquitous average weekly death count model: increasing deaths due to population aging and decreasing deaths due to mortality improvements. As the model adjusts for neither trend, it may be biased in either direction, depending on the relative size of either effect in any given country, although I found a slight tendency towards high excess death estimates across Europe. The lower overall bias when averaging counts instead of rates suggests that errors due to non-adjusted time trends cancel to some degree, demonstrating that an incomplete adjustment to trends can be worse than no adjustment at all.

A major challenge faced by every model is to estimate the magnitude of the future seasonality. In the absence of information about influenza activity, all models just predicted an average seasonality for the following Fall and Winter, possibly with minor temperature adjustments. Whether or not an influenza predictor should be included when estimating excess deaths over the COVID-19 pandemic depends on the purpose of the analysis: causal inference or mortality monitoring. A causal estimate of the direct and indirect effects of the pandemic on overall mortality can not accommodate observed influenza activity past the beginning of the pandemic as this variable is itself influenced by the "treatment", i.e., COVID-19 and associated mitigation measures [33, 34]. In a counterfactual world where the COVID-19 pandemic did not happen, influenza activity throughout 2020 may have been different. For mortality monitoring purposes, however, the inclusion of an influenza predictor is admissible as the question is whether the currently observed death counts are unusual given the currently observed levels of influenza.

How many deaths would we have seen in 2020 if this would have been a typical year? This definition of expected deaths allows to discriminate between suitable and unsuitable models on empirical grounds: We want a model that excels at short-term forecast of weekly death counts. While robustness checks and model averaging are pragmatic way to deal with model specification uncertainty, models of expected deaths are not beyond scrutiny – like other prediction models, they can and should be put to the test. Further research into "short-term mortality prediction" promises a better understanding of systematic biases across different model specifications, which can then

be synthesized into a set of best practices when modeling expected deaths in the absence of unusual events.

Acknowledgments

Supporting information

S1 Fig. Population weighted weekly temperature anomaly by country.

S2 Fig. Cross-validation setup. Rolling origin five-fold cross-validation setup mirroring the task of predicting weekly deaths past the beginning of the COVID pandemic given pre-pandemic data.

S3 Fig. Female weekly and cumulative percent excess death for ages 0 to 65. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S4 Fig. Female weekly and cumulative percent excess death for ages 65 to 75. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S5 Fig. Female weekly and cumulative percent excess death for ages 75 to 85. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S6 Fig. Female weekly and cumulative percent excess death for ages 85 and above. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S7 Fig. Male weekly and cumulative percent excess death for ages 0 to 65. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S8 Fig. Male weekly and cumulative percent excess death for ages 65 to 75. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S9 Fig. Male weekly and cumulative percent excess death for ages 75 to 85. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S10 Fig. Male weekly and cumulative percent excess death for ages 85 and above. Predicted from 9 different models during the year 2020 weeks 8 through 52 for 20 European regions.

S11 Fig. Country ranking of annual excess death percentage. The annual excess death percentage is estimated by age and sex over weeks 8 through 52 of 2020 under 9 different models.

S12 Fig. Percent differences of predicted annual death counts from various models against the 5-year average weekly death count prediction. By age and sex. The prediction differences were summarised across countries by the 0.25, 0.5, and 0.75 quantiles.

S13 Fig. Model-specific bias and error when predicting annual death counts on test data. By age and sex. The error (MAPE, italic) and bias (MPE, bold) measures were summarised across countries by the 0.25, 0.5, and 0.75 quantiles.

References

1. The Economist Data Team. Tracking COVID-19 excess deaths across countries. The Economist. 2020;.

2. Wu J, McCann A. 25,000 Missing Deaths: Tracking the true toll of the coronavirus crisis. New York Times. 2020;.

3. Leon DA, Shkolnikov VM, Smeeth L, Magnus P, Pechholdová M, Jarvis CI. COVID-19: a need for real-time monitoring of weekly excess deaths. The Lancet. 2020;395(10234):e81. doi:10.1016/S0140-6736(20)30933-8.

4. Kontis V, Bennett JE, Rashid T, Parks RM, Pearson-Stuttard J, Guillot M, 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. 2020;doi:10.1038/s41591-020-1112-0.

5. Bilinski A, Emanuel EJ. COVID-19 and Excess All-Cause Mortality in the US and 18 Comparison Countries. JAMA. 2020;324(20):2100. doi:10.1001/jama.2020.20717.

6. Rizzi S, Vaupel JW. Short-term forecasts of expected deaths. Proceedings of the National Academy of Sciences. 2021;118(15):e2025324118. doi:10.1073/pnas.2025324118.

7. Islam N, Shkolnikov VM, Acosta RJ, Klimkin I, Kawachi I, Irizarry RA, et al. Excess deaths associated with COVID-19 pandemic in 2020: age and sex disaggregated time series analysis in 29 high income countries. BMJ. 2021; p. 1–14. doi:10.1136/bmj.n1137.

8. Nørgaard SK, Vestergaard LS, Nielsen J, Richter L, Schmid D, Bustos N, et al. Real-time monitoring shows substantial excess all-cause mortality during second wave of COVID-19 in Europe, October to December 2020. Eurosurveillance. 2021;26(2):1–5. doi:10.2807/1560-7917.ES.2021.26.1.2002023.

9. i Arolas HP, Acosta E, López-Casasnovas G, Lo A, Nicodemo C, Riffe T, et al. Years of life lost to COVID-19 in 81 countries. Scientific Reports. 2021;11(1). doi:10.1038/s41598-021-83040-3.

10. Campbell A, Ward S. Comparisons of all-cause mortality between European countries and regions: 2020. Office for National Statistics; 2021. Available from: https://www.ons.gov.uk/peoplepopulationandcommunity/

birthsdeathsandmarriages/deaths/articles/
comparisonsofallcausemortalitybetweeneuropeancountriesandregions/
2020.

11. Magnani C, Azzolina D, Gallo E, Ferrante D, Gregori D. How Large Was the
    Mortality Increase Directly and Indirectly Caused by the COVID-19 Epidemic?
    An Analysis on All-Causes Mortality Data in Italy. International Journal of
    Environmental Research and Public Health. 2020;17(10):3452.
    doi:10.3390/ijerph17103452.

12. Michelozzi P, de'Donato F, Scortichini M, Pezzotti P, Stafoggia M, Sario MD,
    et al. Temporal dynamics in total excess mortality and COVID-19 deaths in
    Italian cities. BMC Public Health. 2020;20(1). doi:10.1186/s12889-020-09335-8.

13. Modig K, Ahlbom A, Ebeling M. Excess mortality from COVID-19: weekly
    excess death rates by age and sex for Sweden and its most affected region.
    European Journal of Public Health. 2020;31(1):17–22.
    doi:10.1093/eurpub/ckaa218.

14. Serfling RE. Methods for Current Statistical Analysis of Excess
    Pneumonia-Influenza Deaths. Public Health Reports. 1963;78(6):494.
    doi:10.2307/4591848.

15. Barnard S, Fox S, Baker A, Burton P, Goldblatt P, Fitzpatrick J. Excess
    mortality in England. Methodology for the weekly reports. Public Health
    England; 2020. GW-138.

16. Weinberger DM, Chen J, Cohen T, Crawford FW, Mostashari F, Olson D, et al.
    Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the
    United States, March to May 2020. JAMA Internal Medicine.
    2020;doi:10.1001/jamainternmed.2020.3391.

17. Woolf SH, Chapman DA, Sabo RT, Weinberger DM, Hill L, Taylor DDH. Excess
    Deaths From COVID-19 and Other Causes, March-July 2020. JAMA.
    2020;324(15):1562. doi:10.1001/jama.2020.19545.

18. EuroMOMO. EuroMOMO weekly bulletins; 2020. online. Available from: https://www.euromomo.eu/.

19. Aburto JM, Kashyap R, Schöley J, Angus C, Ermisch J, Mills MC, et al. Estimating the burden of the COVID-19 pandemic on mortality, life expectancy and lifespan inequality in England and Wales: a population-level analysis. Journal of Epidemiology and Community Health. 2021;0:1–6. doi:10.1136/jech-2020-215505.

20. Scortichini M, dos Santos RS, Donato FD, Sario MD, Michelozzi P, Davoli M, et al. Excess mortality during the COVID-19 outbreak in Italy: a two-stage interrupted time-series analysis. International Journal of Epidemiology. 2020;49(6):1909–1917. doi:10.1093/ije/dyaa169.

21. Modi C, Böhm V, Ferraro S, Stein G, Seljak U. Total COVID-19 Mortality in Italy: Excess Mortality and Age Dependence through Time-Series Analysis; 2020.

22. Jdanov D, Shkolnikov VM, Galarza AA, Boe C, Barbieri M. Short-Term Mortality Fluctuations Dataseries methods protocol. Max Planck Institute for Demographic Research; 2021. Available from: https://mortality.org/Public/STMF_DOC/STMFNote.pdf.

23. Németh L, Jdanov DA, Shkolnikov VM. An open-sourced, web-based application to analyze weekly excess mortality based on the Short-term Mortality Fluctuations data series. PLOS ONE. 2021;16(2):e0246663. doi:10.1371/journal.pone.0246663.

24. HMD. Human Mortality Database; 2021. Available from: https://mortality.org.

25. Climate Prediction Center. Global Daily Temperature; 2021. Available from: https://psl.noaa.gov/data/gridded/data.cpc.globaltemp.html.

26. Center for International Earth Science Information Network. Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11; 2018.

27. Hager T. Nager.Date – Worldwide Public Holiday; 2021. Available from: https://github.com/nager/Nager.Date.

28. Rue H, Held L. Gaussian Markov random fields. Theory and applications. vol. 104. Chapman & Hall/CRC; 2005.

29. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2009;71(2):319–392. doi:10.1111/j.1467-9868.2008.00700.x.

30. Rue H, Lindgren F, Simpson D, Martino S, Krainski ET, Bakka H, et al.. INLA R library: Full Bayesian Analysis of Latent Gaussian Models using Integrated Nested Laplace Approximations; 2021. Available from: https://www.r-inla.org.

31. Leon DA. Trends in European life expectancy: a salutary view. International Journal of Epidemiology. 2011;40:271–277. doi:10.1093/ije/dyr061.

32. Aburto JM, Schöley J, Kashnitsky I, Zhang L, Rahal C, Missov TI, et al. Quantifying impacts of the COVID-19 pandemic through life expectancy losses: a population-level study of 29 countries. medRxiv. 2021;doi:10.1101/2021.03.02.21252772.

33. Cowling BJ, Ali ST, Ng TWY, Tsang TK, Li JCM, Fong MW, et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. The Lancet Public Health. 2020;5(5):e279–e288. doi:10.1016/S2468-2667(20)30090-6.

34. Soo RJJ, Chiew CJ, Ma S, Pung R, Lee V. Decreased Influenza Incidence under COVID-19 Control Measures, Singapore. Emerging Infectious Diseases. 2020;26(8):1933–1935. doi:10.3201/eid2608.201229.