
# Methods for estimating the excess mortality associated with the COVID-19 pandemic

World Health Organisation

April 5, 2023

# Contents

Glossary

| Term | Explanation |
|---|---|
| All-causes mortality (ACM) | The total deaths (across all causes-of-death) that have occurred in a specified location within a specified period of time e.g. week, month, or year. For country $c$ and time period $t$, represented by $Y_{c,t}$ |
| COVID-19 | Coronavirus disease (COVID-19) is an infectious disease caused by the SARS-CoV-2 virus. |
| Expected deaths | The hypothetical or "counterfactual" total death numbers which for country $c$ and time period $t$ are represented by $E_{c,t}$ and are forecasted using historic (prior to the pandemic) deaths data. |
| Excess deaths | Defined as the difference in death numbers when comparing pandemic ACM, $Y_{c,t}$, to the expected, $E_{c,t}$, which for country $c$ and time period $t$, are calculated as: $$\delta_{c,t} = Y_{c,t} - E_{c,t}$$ |
| Negative excess deaths | When the pandemic ACM for country $c$ over period $t$ is lower than the expected this leads to a mortality deficit or "negative" excess i.e., $$\delta_{c,t} = Y_{c,t} - E_{c,t} < 0$$ Generally, when and where this occurs it is due to declines in death numbers for certain causes during the pandemic period e.g., lower than expected traffic related mortality. However it can be an artifact of the model assumptions e.g., overestimate of expected or underestimate of ACM. |

# 1 Introduction

The World Health Organisation (WHO) has been tracking the impact of COVID-19 throughout the course of the pandemic. Although data on the number of cases and deaths are being reported to WHO and made publicly available at https://covid19.who.int/, these figures do not offer a comprehensive picture of the health burden attributable to COVID-19, or the total number of lives lost as a result of the pandemic. Some deaths attributable to COVID-19 may not have been officially certified as such, as testing may not have been conducted before the individual's passing. In addition, countries have varied in their application of death certification protocols related to COVID-19 (Riffe and Acosta, 2021).

The impact of the pandemic have extended far beyond the number of deaths directly caused by COVID-19. On one hand, the conditions that arose during the pandemic, such as overwhelmed health systems or patients avoiding healthcare, have led to additional deaths. On the other hand, in countries where COVID-19 spread was limited due to lockdown measures and other non-pharmaceutical interventions, there have been decreases in potential causes of death, such as those attributable to air pollution, traffic accidents, or other communicable diseases like influenza, which have resulted in negative excess or deficit deaths (Kung *et al.*, 2020; Karlinsky and Kobak, 2021).

In light of the challenges posed by using reported COVID-19 data, excess mortality is considered a more objective and comparable measure of the mortality impact of COVID-19 across countries (Leon *et al.*, 2020). The WHO defines excess mortality as, "the mortality above what would be expected based on the non-crisis mortality rate in the population of interest" (Checchi and Roberts, 2005). Understanding excess mortality not only provides a more comprehensive view of the pandemic's impact, but can also help in the development and implementation of effective public health initiatives.

The all-cause mortality (ACM) counts in a given country $c$ and month $t$ during 2020 and 2021 are represented as $Y_{c,t}$. These counts, in addition to the contribution from expected deaths, are assumed to be a result of the direct effects of COVID-19 (i.e., deaths attributable to it) and the indirect knock-on effects on health systems and society, along with deaths that were averted. Using a monthly time scale provides sufficient temporal resolution for most public health purposes. The hypothetical or "counterfactual" no-COVID-19 scenario is represented as the expected death numbers $E_{c,t}$ forecasted to the same month $t$, using historic deaths data prior to the pandemic.

Excess deaths are defined as:

$$\delta_{c,t} = Y_{c,t} - E_{c,t} \qquad (1)$$

for country $c$ where $c = 1,\ldots,194$, and in month $t$ where $t = 1,\ldots,24$ represent months in 2020 and 2021.

Estimating excess deaths for all countries is a challenging task, as the required ACM counts $Y_{c,t}$ are currently unavailable for many country/month combinations. Routine mortality data is typically reported to the WHO a year or more after the year of death. Furthermore, differences in reporting capacities and data quality across countries mean that many nations lack the infrastructure to provide reliable routine data even historically (Mikkelsen *et al.*, 2015; Adair and Lopez, 2018; GBD, 2020; UNSD, 2021; Karlinsky, 2021). As a result, these countries were unable to monitor ACM during the unprecedented COVID-19 pandemic, making it difficult to contribute to the centralized systematic mortality surveillance necessary to measure global, regional and country level excess mortality by the WHO.

This report details the ongoing methods development to generate WHO excess mortality estimates. Section 2 outlines the data sources used, while section 3 describes the models used to estimate the expected death numbers. Section 4 describes our national covariate model and Section 5 outlines the models used for countries with subnational monthly data, national annual data, or a combination of both. The report also includes a description of the current approach used to derive preliminary age- and sex-distributions of excess deaths, which is covered in section 6.

# 2 Data Sources

## 2.1 Mortality Data

Excess mortality cannot be directly measured for all countries due to many not having the required ACM data. The WHO typically receives routine mortality data on an annual basis in the year after the year of death or after an even greater lag. Civil registration and vital statistics (CRVS) systems differ greatly across countries with varying timelines and quality control measures for compiling unit record cause-of-death numbers into aggregates identified by cause, age, sex, place, and period of death. In addition, differential reporting coverage, the absence of electronic surveillance systems in some locations and limited investments in CRVS systems has resulted in many nations lacking the structures necessary to provide good quality routine data, even before the COVID-19 pandemic. This lack of capacity and data required to monitor ACM has been further exacerbated during the unprecedented pandemic, rendering many countries unable to contribute to a centralized systematic mortality surveillance that would be needed to measure global, regional and country level excess mortality by the WHO.

All countries report their official COVID-19 death count, but for many countries it is anticipated that the official counts may not be accurate, with serious underestimation, due to the reasons previously mentioned and for political reasons. Nonetheless, the official counts are used as the best available data for comparison with the estimated excess deaths, and the COVID-19 death rate is used as a covariate in the ACM estimation model.

The primary sources of data for this study are reports of ACM collected and reported by relevant institutions in each country, including national statistics offices, ministries of health and population registries. These data have been compiled from various repositories such as the data routinely shared with WHO as part of its standing agreement with member states, Eurostat, The Human Mortality Database (HMD) as part of the Short-Term Mortality Fluctuations (STMF) project (Németh *et al.*, 2021) and the World Mortality Dataset (WMD), as described in Karlinsky and Kobak (2021). Monthly data are included after accounting for delayed registration either by adjusting for registration delay (Australia, Brazil, United States) or by not-including highly incomplete months.

Table 1 provides a summary of available data for the 194 WHO Member States. The ”Full national” group includes countries that have data for all 24 months between January 2020 and December 2021; The ”Partial national” group include countries that have data for less than 24 months; The table also includes a category for countries with ”subnational and/or annual data” and those without any available ACM data accessible to

the WHO. Groupings in the table are according to the WHO regions: African Region (AFR), Region of the Americas (AMR), Eastern Mediterranean Region (EMR), European Region (EUR), South-East Asia Region (SEAR), Western Pacific Region (WPR).

Table 1: *Data availability summary for 2020 and 2021 for WHO Member States.*

| Region | Full National | Partial National | Subnational and/or Annual | No Data | Total |
|---|---|---|---|---|---|
| AFR | 3 | 2 | 0 | 42 | 47 |
| AMR | 23 | 4 | 1 | 7 | 35 |
| EMR | 8 | 2 | 0 | 11 | 21 |
| EUR | 51 | 1 | 0 | 1 | 53 |
| SEAR | 2 | 1 | 2 | 6 | 11 |
| WPR | 9 | 0 | 2 | 16 | 27 |
| Global | 96 | 10 | 5 | 83 | 194 |

The current report reflects the current state of data availability, which is expected to improve over time. As shown in Table 1, just over a half (106) of the 194 Member States provide monthly national data for at least part of the pandemic period, while 5 other countries provide subnational monthly data, national annual data, or a combination of the two. However, there is a significant regional disparity in data availability, with EUR being almost fully represented, AMR having data from 80% of the countries, and other regions being more poorly represented. For instance, in AFR, only 5 out of 47 countries have data.

As mortality data may be incomplete due to various reasons, such as incomplete registration or underreporting, adjusting reported deaths for completeness is needed for countries with available data. Adjusting for completeness is a demographic method used to estimate the total number of deaths that occurred in a population including those that may not have been recorded in the official mortality data. Details can be found in Global Health Estimates method's document (ghe2019_life-table-methods.pdf (who.int)). Reported deaths are adjusted as follows:

$$Y_{c,t} = \frac{Y'_{c,t}}{completeness_{c,t}},$$

where $Y'_{c,t}$ represents the non-adjusted reported deaths.

## 2.2 Covariate Data

For countries with no data, we predict the ACM count using a log-linear covariate model. A range of covariates were considered, including those measured in the pandemic period i.e.,

- a high income country binary indicator,
- the COVID-19 test positivity rate,
- the COVID-19 death rate,
- temperature,
- population density,
- a socio-demographic index (SDI) and the human development index (HDI),
- a measure of stringency (index for lockdown restrictions and closures, overall government response),
- economic measures (including measures such as income support and debt relief),
- containment[1],

and those representing pre-pandemic (for the year 2019) country characteristics:

- non-communicable disease rate,
- cardiovascular disease rate,
- HIV death rate,
- diabetes prevalence rate,
- life expectancy at birth,
- proportion of the population under-15
- proportion of the population over-65.

Some of the covariates are time-varying (COVID-19 test positivity rate, COVID-19 death rate, temperature, stringency, overall government response, containment), while the remainder are constant over time. A number of the covariates were not available for all months for all countries and so missing values were imputed using WHO regional medians. The final set of covariates was chosen based on the lowest out-of-sample error.

[1]The containment measure combines "lockdown" restrictions and closures with measures such as testing policy and contact tracing, short term investment in healthcare, as well investments in vaccines – it is calculated using all ordinal containment and closure policy indicators and health system policy indicators, for further details see Hale *et al.* (2020)

# 3 Deriving expected mortality for years 2020 and 2021

A key component of the excess mortality calculation is the ACM count that would be expected in non-pandemic times, for each country and month. We describe models for two types of countries: those that have historic *monthly* ACM data, and those that have historic *annual* ACM data only. In terms of the period upon which we base the expected numbers, it is usually 2015–2019 for countries with monthly historical data, and it is usually 2000–2019 for countries with annual historical data.

## 3.1 Countries with Monthly Data

We consider first those countries with monthly ACM data over multiple years. For country $c$, $Y_{c,t}$ represents the ACM count for country $c$ and month $t$, for $t = 1,\ldots,M_c$, where $M_c$ is the number of historic months for which we have data. We assume the sampling model for $Y_{c,t}$ is negative binomial:

$$Y_{c,t}|\mu_{c,t} \sim \text{NegBin}\left(\mu_{c,t}^{\text{E}}, \phi_c^{\text{E}}\right),$$

parametrized in terms of the mean, $\mu_{c,t}^{\text{E}}$, and the overdispersion parameter, $\phi_c^{\text{E}}$, such that $\text{var}\left(Y_{c,t}|\mu_{c,t}^{\text{E}}, \phi_c^{\text{E}}\right) = \mu_{c,t}^{\text{E}}\left(1 + \mu_{c,t}^{\text{E}} / \phi_c^{\text{E}}\right)$, with the Poisson model being recovered as $\phi_c^{\text{E}} \to \infty$. We let $v[t]$ index the year in which month $t$ occurred (for example, labeled 1, . . . , 5 when data are available for 2015–2019) and $m[t]$ be the month (labeled 1, . . . , 12), so that given $v, m$ we can find $t$ as $t = 12(v - 1) + m$. The mean is modeled via the log-linear form,

$$\eta_{c,t} = log\left(\mu_{c,t}\right) = f_c^{y}(v[t]) + f_c^{m}(m[t]) \qquad (2)$$

where $f_c^{\text{y}}(\cdot)$ models the *annual trend*, and $f_c^{\text{m}}(\cdot)$ is a smooth function of time t which accounts for *within-year* seasonal variation. The yearly trend is modeled with a linear model and within-year variation with a cyclic cubic spline (Rivera et al., 2020). For the latter, we use the `gam` function in the `mgvc` package with REML used to select smoothing parameters (and with the default settings). The models are fitted separately for each country and used to obtain predictions of the expected deaths $\mu_{c,t}^{\text{E}}$ for all $t$ in 2020 and 2021, with both a point estimate and a standard deviation being produced, and these can be viewed as summaries of the posterior distribution, see Section 6.10 of Wood (2017) for details.

## 3.2 Countries with Annual Data

For countries with only annual historic data, the goal is to predict expected numbers by month $t$ for $t = 1,\ldots,24$. We summarize our strategy for producing expected numbers for countries with annual data only, before giving details:

1. Fit a negative binomial linear model to the countries with annual counts only. Use the spline to predict the total annual ACM for 2020 and 2021, for these countries.
2. In a separate exercise, fit the multinomial model to all of the countries with monthly data, with deaths being attributed via a log-linear temperature model.
3. Combine the linear model with the multinomial model using monthly temperature apportionment to obtain expected numbers for the countries without monthly data.

The annual trend can be estimated for each country using the method we described in the previous section minus the monthly term, i.e., by using a linear model for year. To apportion the yearly totals to the months, we use the fact that a collection of Poisson random variables conditioned on their sum produce a multinomial distribution with within-year variation modeled using temperature, which is acting as a surrogate for seasonality. It is well-known that mortality is associated with temperature (see for example Parks *et al.* (2018)), and we wanted a relatively simple model, using a well-measured variable. This relationship is learned from countries with historic monthly data. We use a smooth series of monthly temperatures since 2015. Let $\boldsymbol{Y}_{c,v} = \{Y_{c,v,m}, m = 1, \ldots, 12\}$ be the vector that contains the ACM counts by month in year $v$, $v = 1,\ldots,5$, for country $c$. Suppose each of the 12 constituent counts are Poisson with mean $\zeta_{c,v,m}$, for $m = 1,\ldots,12$. Then, within year $v$, conditional on the total ACM, $Y^{+}_{c,v}$,

$$\boldsymbol{Y}_{c,v} | Y^{+}_{c,v}, \boldsymbol{p}_{c,v} \sim \text{Multinomial}\left(Y^{+}_{c,v}, \boldsymbol{p}_{c,m}\right),$$

where $\boldsymbol{p}_{c,v} = \{p_{c,v,m}, m = 1, \ldots, 12\}$ with,

$$p_{c,v,m} = \frac{\zeta_{c,v,m}}{\sum_{m=1}^{12} \zeta_{c,v,m}},$$

We assume,

$$\log(\zeta_{c,v,m}) = z_{c,v,m}\beta \qquad (3)$$

where $z_{c,v,m}$ is the temperature and $\beta$ is the associated log-linear coefficient; no intercept is needed in the log-linear model, since when we take the ratio, to form the multinomial prob- abilities, if included, it would cancel. The multinomial model can be fitted in INLA using the Poisson trick (Baker, 1994) which involves fitting the Poisson model for the data in country $c$, month $m$:

$$Y_{c,v,m}|\lambda_{c,v} \sim \text{Poisson}(\lambda_{c,v}e^{z_{c,v,m}\beta}),$$

where the $\lambda_{c,v}$ parameters are given (improper) priors $\pi(\lambda_{c,v}) \propto 1/\lambda_{c,v}$. We use the default INLA prior for $\beta$, which is a normal with a large variance. Further details of the Poisson trick may be found in Section 3 of the Supplementary Material (Knutson *et al.*, 2023).

### 3.3 Modeling Uncertainty in the Expected Numbers

For all countries the expected numbers appear directly in the excess calculation, (1). In addition, for countries with no pandemic ACM data, the Poisson model we adopt for covariate modeling includes the expected number as an offset. For all countries and months, we obtain not just an estimate of the mean expected mortality but also a measure of the uncertainty (due to uncertainty in estimating the linear + spline model) in this estimate. We now describe how the uncertainty in the mean expected count is accounted for in our modeling.

For countries with monthly data, we use the annual linear and monthly spline model to predict the log of the mean expected number of deaths. Asymptotically, the estimator for the log of the mean expected numbers is normally distributed. Let $\hat{\eta}_{c,t'}$ and $\hat{\sigma}^2_{c,t'}$ represent the mean and standard deviation of the log prediction for pandemic months, labeled as $t' = 1, \ldots, 24$. We stipulate $S$ samples from the asymptotic normal sampling distribution with mean $\hat{\eta}_{c,t'}$ and standard deviation $\hat{\sigma}_{c,t'}$; denote these samples by $\eta^{(s)}_{c,t'}, s = 1, \ldots, S$. We then transform the samples so that we have samples for the expected numbers $E^{(s)}_{c,t'} = \exp\left(\eta^{(s)}_{c,t'}\right)$, for $s = 1, \ldots, S$. We then use the method of moments to fit a gamma distribution of these $S$ samples with shape $\tau_{c,t}$ and rate $\tau_{c,t'}/E_{c,t'}$. In particular, letting $m_{c,t'}$ denote the sample mean, and $V_{c,t'}$ denote the sample variance, we set $\hat{E}_{c,t'} = m_{c,t'}$ and $\hat{\tau}_{c,t'} = \frac{m^2_{c,t'}}{V_{c,t'}}$. We approximate the distribution of the expected numbers as gamma since this is conjugate to the Poisson, and so allows efficient inference with INLA (Rue *et al.*, 2009) using a negative binomial model, as we describe in Section 4. Effectively, we are approximating the sampling distribution of the mean expected count by a gamma distribution.

We now consider a generic country $c$ with yearly data only. In pandemic year $v'$, we use the linear model to predict the log of the expected number of deaths. Let $\hat{\eta}_{c,v'}$ and $\hat{\sigma}^2_{c,t'}$ represent the mean and standard deviation of the prediction, for (the two pandemic years). We then simulate $S$ samples from a normal distribution with mean $\hat{\eta}_{c,v'}$ and standard deviation $\hat{\sigma}_{c,v'}$; denote these samples by $\eta^{(s)}_{c,v'}$, $s = 1,\ldots,S$. We then transform the samples so that we have samples for the expected numbers $E^{(s)}_{c,v'} = \exp\left(\eta^{(s)}_{c,v'}\right)$, for $s = 1, \ldots, S$.

We then apply the monthly temperature model to produce predictions of the proportion of deaths in each month in each year, i.e., for a given pandemic month $m'$, we have $S$ samples of the predicted proportion of deaths in month $m'$ of year $v'$, $p^{(s)}_{c,v',m'}$, for $s = 1,\ldots,S$. Converting to pandemic months $t' = 12(v' - 1) + m'$ we then produce samples of the expected number of deaths in month $t'$, as

$$E^{(s)}_{c,t'} = E^{(s)}_{c,v'} \times p^{(s)}_{c,v',m'}.$$

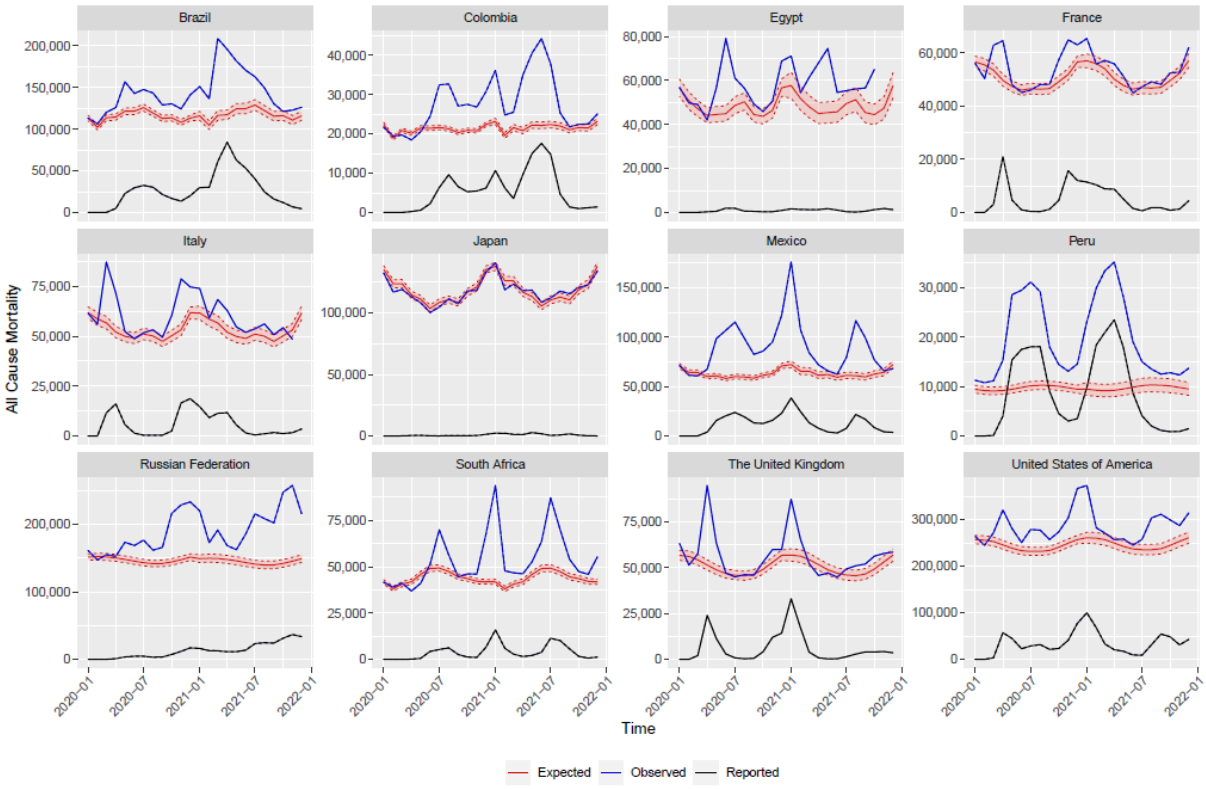
We then use the method of moments to fit a gamma distribution to these $S$ samples, as for the countries with monthly data. To summarize, in both cases we have a distribution for $E_{c,t'}$which is $\text{Gamma}\left(\hat{\tau}_{c,t'}.\hat{\tau}_{c,t'}/\hat{E}_{c,t'}\right)$. In Section 3 of the Supplementary Materials (Knutson *et al.*, 2023) we provide comparisons of the true distribution of the mean expected counts and the approximating gamma distributions, and illustrates that the latter are accurate. We also experimented with including negative binomial sampling variability in the calculation of the expected numbers, but it made little additional contribution to the intervals for the excess.

In the next section we describe a Bayesian model for the ACM counts in the pandemic, for countries without data. As we have described above, inference for the expected numbers is an approximation to a Bayesian analysis. We sample from the asymptotic normal distribution of the prediction estimator which will approximate a Bayesian analysis with (improper) flat priors. Hence, when we combine the two components in the excess (1) we view the resultant inference as Bayesian. We next describe how we model ACM – we have different models for different data scenarios but in each case the starting point is the Poisson distribution.

# 4 National Mortality Models for Countries with No Data

For countries with observed monthly national ACM data, $Y_{c,t'}$, we use these directly in the excess calculation. For the countries with no data we need to estimate the ACM count. We follow a Bayesian approach so that for countries without data we obtain a predictive distribution over this count and this, when combined with the gamma distribution for the expected numbers, gives a distribution for the excess $\delta_{c,t}$.

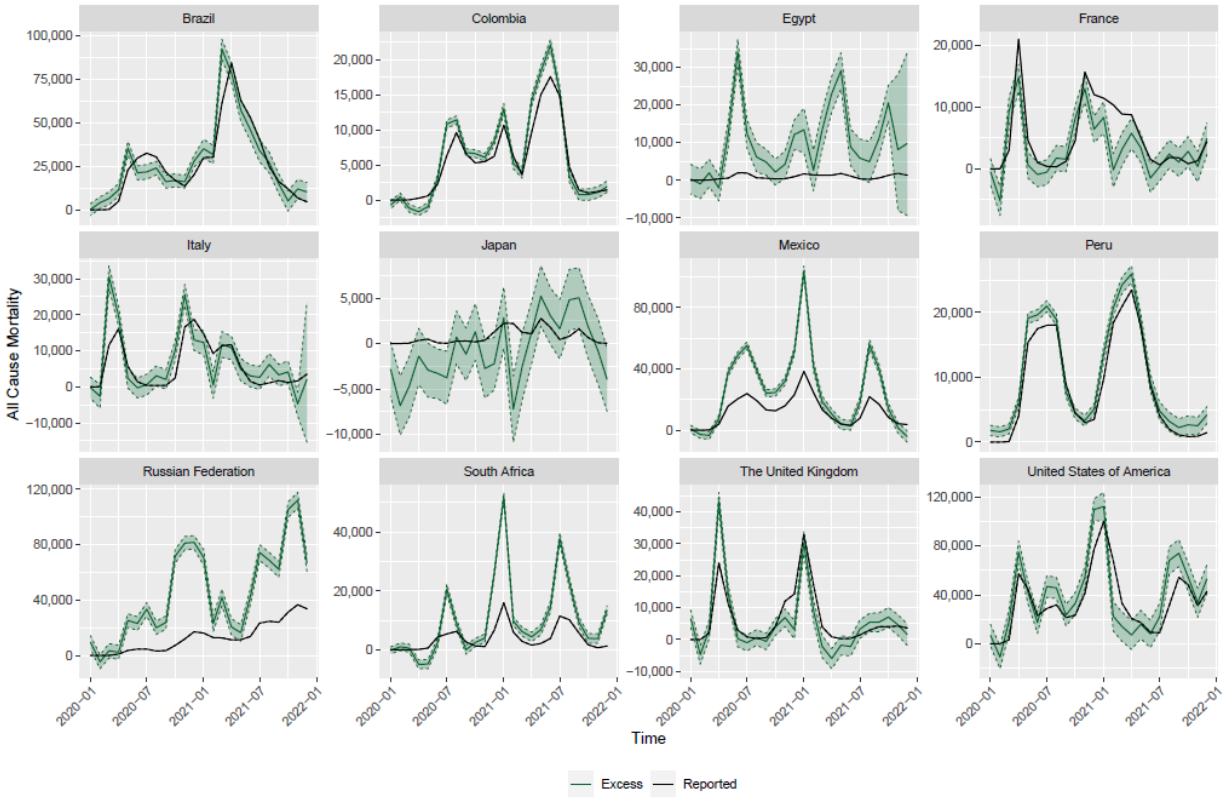
Figure 1: *Monthly time series of ACM counts, expected counts (with 95% interval estimates) and reported COVID-19 mortality counts, for selected countries. ACM counts are available for all months apart from Egypt, for which the last month is missing.*




In Figure 1 we plot the monthly counts for a range of countries with monthly ACM data, along with the reported COVID-19 deaths and the expected numbers. We see very different scenarios in different countries. In all countries but Japan there is a clear large difference between the observed and the expected, though within each country this difference shows large fluctuations over time. In Figure 2, again for countries with monthly ACM data, we plot the excess $\delta_{c,t} = Y_{c,t} - E_{c,t}$ as a function of month $t$ (including uncertainty in the expected numbers), along with the reported COVID-19 deaths. As expected, $\delta_{c,t}$ is greater than the reported overall in general, except in Japan, and for most countries displayed the difference between the excess and the reported

shows a complex temporal pattern.

Figure 2: *Monthly time series of excess mortality, along with reported COVID-19 mortality counts. ACM counts are available for all months apart from Egypt, for which the last month is missing. For this month, the covariate prediction model is used for the point and interval estimates.*




While complex models that attempt to pick up data nuances are desirable, given the idiosyncrasies of the different data sources described in Section 2, any modeling exercise is fraught with difficulties, and we resort to a relatively simple model in which we build an overdispersed Poisson log-linear regression model for the available monthly ACM data to predict the monthly ACM in those countries with no data. We cannot overemphasize the regional im- balance of the missing ACM data – in the AFR region in particular, our estimates should be viewed with extreme caution, since they are predicted from data which overwhelmingly is from other regions.

The basic starting model is

$$Y_{c,t}|E_{c,t},\theta_{c,t} \sim \text{Poisson}\left(E_{c,t}\theta_{c,t}\right) \qquad (4)$$

so that $\theta c,t > 0$ is a relative rate parameter, with the cases $\theta c,t > 1$ and $\theta c,t < 1$ corresponding to higher and lower ACM rates than expected, based on historic data. Recall, from Section 3, that we model the distribution of the expected counts $E_{c,t}$ as $\text{Gamma}(\hat{\tau}_{c,t}, \hat{\tau}_{c,t}/\hat{E}_{c,t})$.

When combined with (4), we obtain the sampling model,

$$Y_{c,t}|\theta_{c,t} \sim \text{NegBin}(\hat{E}_{c,t}\theta_{c,t}, \hat{\tau}_{c,t})$$

with known overdispersion parameter $\hat{\tau}_{c,t}$to give $\text{var}(Y_{c,t}|\theta_{c,t}) = \hat{E}_{c,t}\theta_{c,t}(1 + \hat{E}_{c,t}\theta_{c,t}/\hat{\tau}_{c,t})$. The mean is $\text{E}[Y_{c,t}|\theta_{c,t} = \hat{E}_{c,t}\theta_{c,t}$ and the relative rate parameter $\theta_{c,t}$ is modeled as,

$$\log\theta_{c,t} = \alpha + \sum_{b=1}^{B}\beta_{bt}X_{bct} + \sum_{g=1}^{G}\gamma_g Z_{gc} + \epsilon_{c,t} \qquad (5)$$

We now describe parameter interpretation and model details.

- The intercept is $\alpha$.
- The time-invariant covariates (e.g., income level, historic cardiovascular and diabetes rates) have fixed log-linear association parameters $\gamma_g$ .
- We have $B$ time-varying covariates (e.g., sqrt(C19 death rate), test positivity rate, containment, temperature, and their interaction with income level) as well as the time invariant covariates, and we allow the log-linear associations for these variables, $\beta_{bt}$, to be time-varying via a random walk of order 2 (RW2) prior (Rue and Held, 2005) which has variance $\sigma_\beta^2$. These parameters include a sum-to-zero constraint, since we include a fixed effect for the overall association (across months) – these are included in the $G$ time-invariant part of the model.
- There are two sources of excess-Poisson variation in our model. The negative binomial component, with known $\hat{\tau}_{c,t}$, arises because of the uncertainty in the expected numbers.
- The Bayesian model is completed by prior specifications on the regression coefficients of the log-linear model and any hyperparameters. We use default priors (normal with large variance) on the intercept and fixed association parameters, and penalized complexity (PC) priors (Simpson et al., 2017) on the RW2 standard deviations and on $\sigma_\epsilon$. Specifically, letting σβ denote a generic RW2 standard deviation parameter, the PC priors are such that $\Pr(\sigma_\beta > 1) = 0.01$, and the PC prior on the overdispersion parameter $\sigma_\epsilon$ has $\Pr(\sigma_\epsilon > 1) = 0.01$.

Each country will clearly have its own specific temporally correlated baseline, as a result of unobserved covariates and model misspecification, but we did not include terms to model such a baseline (using a RW2 or a spline, for example), since fits from this model are not being used to estimate the excess for countries with data. Rather, we are using this model to predict the ACM for countries with no data. Hence, we did not use RW2 intercepts as these would dilute the covariate effects, due to confounding by time (Kelsall *et al.*, 1999), and it is these covariate effects that are key to prediction for countries with no data. If we had included a RW2 baseline, then a country-specific RW2 model would give estimated contributions of zero in countries with no data and so would not provide any benefit. This is but one of the model assumptions that are forced

upon us by the limited data we have available. The country-level model was fitted using the INLA method (Rue *et al.*, 2009) and accompanying R implementation.

For countries with no ACM data, we obtain a predictive distribution by averaging the negative binomial model with respect to the posterior via,

$$\Pr(Y_{c,t}|\boldsymbol{y}) = \int \underbrace{\Pr(Y_{c,t}|\theta_{c,t})}_{\text{Negative Binomial}} \times \underbrace{p(\theta_{c,t}|\boldsymbol{y})}_{\text{Posterior}}\, d\theta_{c,t}.$$

We use INLA to fit the covariate model, and then use the posterior sampling feature to produce samples for the components of (5), which in turn produces samples $\theta_{c,t}^{(s)} \sim p(\theta_{c,t}|\boldsymbol{y})$from the posterior. We then simulate $Y_{c,t}^{(s)}|\theta_{c,t}^{(s)}$ from the negative binomial, for $s = 1, \dots, S$.

Partial monthly data is available for 10 countries, and for these we require a switch from observed data to the covariate modeled ACM. The naive application of the covariate model will lead to the possibility of unrealistic jumps (up or down) when we switch from the observed data to the covariate model, and to alleviate this problem we benchmark the predictions to the last observed data point. We let $T_c^{(1)}$ represent the number of observed months of data and $T_c^{(2)}$ be the number of months for which there are no data, for country c. For a country with partial data, let $\boldsymbol{y}_c^{(1)} = \left[y_{c,1}, \dots, y_{c,T_c^{(1)}}\right]$ represent the observed partial data. We then wish to predict the $\boldsymbol{y}_c^{(2)} = \left[y_{c,T_c^{(1)}+1}, \dots, y_{c,T_c^{(1)}+T_c^{(2)}}\right]$ for the missing period. The model for the missing data period is,

$$y_{c,t}^{(2)}|\boldsymbol{y}_c^{(2)}, \theta_{c,t}, f_c \sim \text{NegBin}\left(\hat{E}_{c,t}\theta_{c,t}f_c, \hat{\tau}_{c,t}\right) \qquad (6)$$

for $t = T_c^{(1)} + 1, \dots, T_c^{(1)} + T_c^{(2)}$, where $\theta_{c,t}$ is a function of the covariates in the missing data period (specifically given by (5)), and the benchmarking factor is,

$$f_c = f_c\left(\theta_{c,T_c^{(1)}}\right) = \frac{y_{c,T_c^{(1)}}}{\hat{E}_{c,T_c^{(1)}}\theta_{c,T_c^{(1)}}},$$

where $\theta_{c,T_c^{(1)}}$ is given by equation (5). This factor matches the last observed death count to the covariate model projected back to the last observed count. This factor is applied subsequently to all of the missing data months. To implement the benchmark, samples from the posteriors for $\theta_{c,t}$ and $f_c$ are used in (6), and then negative binomial counts are drawn.

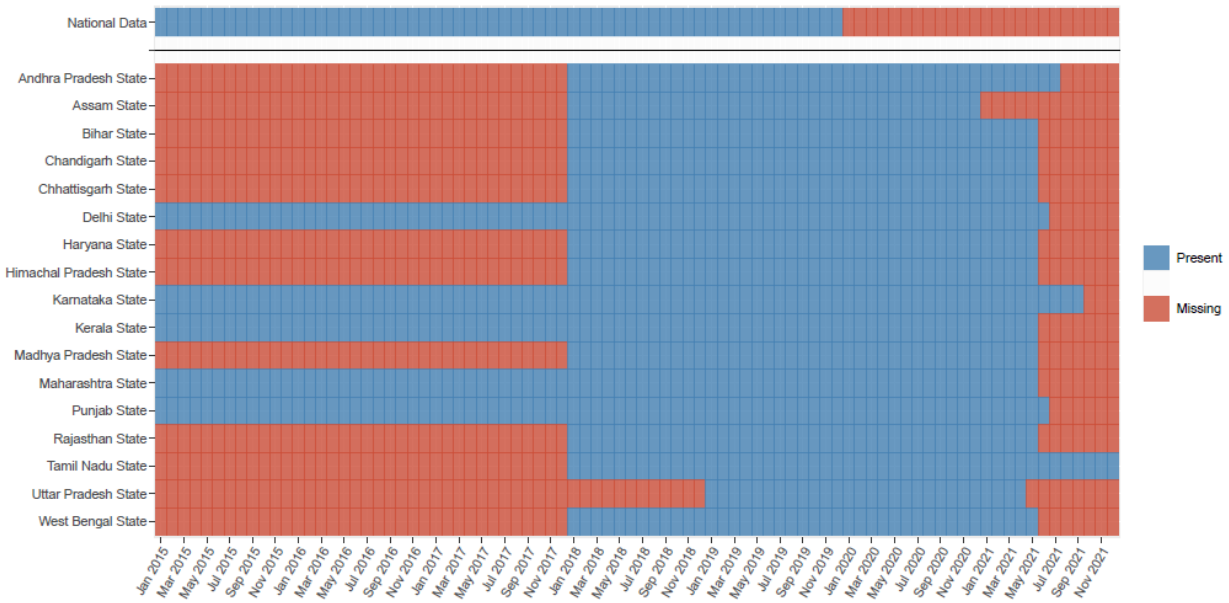
# 5 Observed Mortality Subnational and Annual Data Modeling

For a small number of countries for which national ACM data are not available (India and Indonesia) we instead have ACM data from subregions, with the number of regions with data potentially changing over time. For other countries we obtain national annual ACM data only. In this section we describe the models we use in these situations. For the subnational scenario we construct a statistical model building on, and expanding, a method previously proposed by Karlinsky (2022) that is based on a proportionality assumption.

## 5.1 Subnational Data Model

For Indonesia we have monthly subnational data for 2020 and for the first six month of 2021. India has data from 17 states and union territories over the pandemic period (out of 36), but this number varies by month.

Figure 3: *Plot of missingness in subnational data for India across 2015–2021.*




We consider the most complex subnational scenario in which the number of regions with monthly data varies by month, using India as an example. For India, we use a variety of sources for registered number of deaths at the state and union-territory level. The information was either reported directly by the states through official reports and automatic vital registration, or by journalists who obtained death registration information through Right To Information requests (see Section 6 of the Supplementary Materials (Knutson *et al.*, 2023) for full details).

The available data we have for India is summarized in Figure 3. We assume in total that there are K regions that contribute data at any time. We develop the model for a generic country and hence drop the $c$ subscript.

For the historic data in month $t$ we have total deaths counts along with counts over regions, $Y_{t,k}$, $k \in K_t$, so that in period $t$, $|K_t|$is the number of regions that provide data with $k \in K_t$ being the indices of these areas from $\{1,\ldots,K\}$. We let region 0 denote all other regions, which are not observed in pandemic times, at time $t$ and let $S_t = \{0, K_t\}$. We assume, in month $t$:

$$T_{t,k}|\lambda_{t,k} \sim \text{Poisson}\big(N_{t,k}\lambda_{t,k}\big), \quad k \in S_t,,$$

where $N_{t,k}$ is the population size, and $\lambda_{t,k}$ is the rate of mortality. Hence, the national total in year $t$ is distributed as

$$Y_{t+}|\lambda_{t,k}, k \in S_t \sim \text{Poisson}\left(\sum_{k\in S_t} N_{t,k}\lambda_{t,k}\right).$$

If we condition on the total deaths, we obtain,

$$\boldsymbol{Y}_t|\boldsymbol{p}_t \sim \text{Multinomial}_{|S_t|}(Y_{t+}, \boldsymbol{p}_t),$$

with $\boldsymbol{p_t} = \{p_{t,k}, k \in S_t\}$, and,

$$p_{t,k} = \Pr(\text{death in region } k \mid \text{month } t, \text{death}) = \frac{N_{t,k}\lambda_{t,k}}{N_{t,+}\lambda_{t,+}},$$

Our method hinges on this ratio being approximately constant over time. If, over all regions in which data are observed combined, there are significant changes in the proportions of deaths in the regions as compared to the national total, or changes in the populations within the regions over time, then the approach will be imprecise for estimation of a natural total. However, with multiple regions, we gain robustness since it is the cumulative departure from the constant fractions that is relevant. For India, the fractions of the total ACM by state are shown in Figure 4. There are certainly deviations from constancy for some states, but in general the assumption appears tenable, at least in pre-pandemic periods. Of course, the great unknown is whether the assumption remains reasonable over the pandemic. To address this, we carry out extensive sensitivity and cross-validation analyses (reported in Section 6 of the Supplementary Materials (Knutson *et al.*, 2023)).

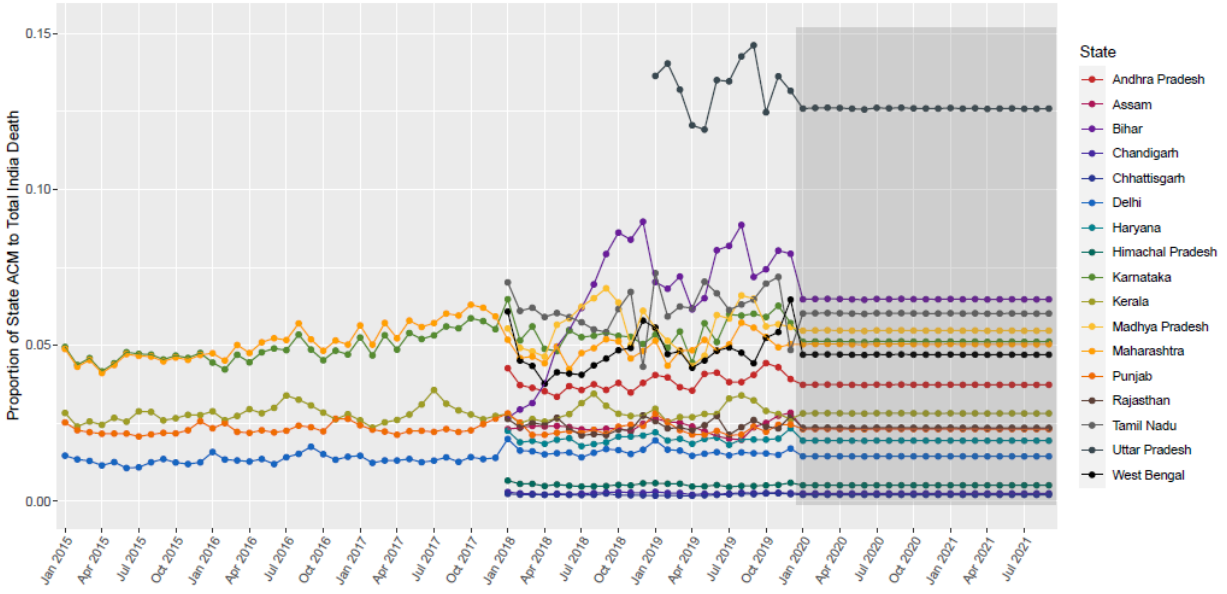
We model the monthly probabilities as,

$$\log\left(\frac{p_{t,k}}{p_{t,K_t+1}}\right) = \alpha_k + e_t, \quad k \in S_t, \qquad (7)$$

where the $\alpha_k$ parameters are unrestricted and $e_t \sim \text{N}(0, \sigma_\epsilon^2)$, and we can examine the size and temporal structure of the error terms $e_t$, to assess the proportionality assumption, at least over the available pre-pandemic period. We emphasize that we

do not use any covariates in the subnational model, but infer the national ACM from the subnational contributions.

Figure 4: *Plot of estimated proportion of subnational deaths to national deaths by state in pre-pandemic and pandemic periods (grey rectangle). The horizontal flat lines are the point estimates for the fraction for the respective states during the pandemic months.*




To specify the model, we take a multinomial with a total number of categories that corresponds to all regions that appear in the data, $K$, along with a final category for the unknown remainder. We specify the likelihood over all months by exploiting the property that a multinomial collapsed over cells is also multinomial. Hence, in year t we have a multinomial with $|K_t| + 1$ categories with constituent probabilities constructed from the full set of $K + 1$ probabilities.

To derive the predictive distribution, we abuse notation and let $Y_{t,1}$ denote the total number of observed subnational deaths at time $t$, and $Y_{t,2}$ the total number of unobserved subnational deaths at time $t$, with $Y_{t,+} = Y_{t,1} + Y_{t,2}$ being the total (national) number of deaths at time $t$. Hence, at time $t$, $Y_{t,1}|p_t, Y_{t,+} \sim \text{Binomial}(Y_{t,+}, p_t)$, where $p_t = \sum_{k \in K_t} p_{t,k}$. In order to fit the multinomial model in a Bayesian framework and predict the total number of deaths in 2020– 2021, we need to specify a prior for $Y_{t,2}$ or, equivalently, for $Y_{t,+}$, where $t$ indexes months in this period. We will use the prior $p(Y_{t,+}) \propto 1/Y_{t,+}$, which is a common non-informative prior for a binomial sample size (Link, 2013), and has the desirable property that the posterior mean for $Y_{t,2}$, conditional on $p_t$, is $\mathrm{E}[Y_{t,2}|p_t] = Y_{t1}(1 - p_t)/p_t$, i.e., of the same form as the simple frequentist "obvious" estimator, which leads to the naive estimate of the ACM, $Y_{t,1} + \hat{Y}_{t,2} = Y_{t,1}/p_t$.

To give more details for implementation we will use a general result. Suppose

$$Y_{t,1}|Y_{t,+}, p_t \sim \text{Binomial}\left(Y_{t,+}, p_t\right)$$
$$p\left(Y_{t,+}\right) \propto 1/Y_{t,+},$$

so that, in particular, the marginal distribution of $Y_{+t}$ does not depend on $p_t$. Then the posterior for the missing ACM count, conditional on $p_t$, is

$$Y_{t,+}|Y_{t,1}, p_t \sim Y_{t,1} + \text{NegBin}\left(Y_{t,+}, 1 - p_t\right)$$

or, equivalently,

$$Y_{t,+} - Y_{t,1}|Y_{t,1}, p_t \sim \text{NegBin}\left(Y_{t,1}, 1 - p_t\right).$$

This links to one of the usual motivations for a negative binomial (number of trials until we observe a certain fixed number of events) — making inference for the number of total deaths it takes to produce $Y_{t,1}$ deaths in the sub-regions. We implement this model in Stan. In Section 4 of the Supplementary Materials (Knutson *et al.*, 2023) we detail a simulation study that validates the method in the situation in which the missing data follow the assumed form.

For the other countries with subnational data, the number of subregions is constant over time, and so in the above formulation the multinomial is replaced by a binomial. Details for these countries are in Section 6 of the Supplementary Materials (Knutson *et al.*, 2023). For Indonesia we have subnational data from only Jakarta at the monthly level and historic national ACM at the annual level. Hence, we fit a binomial subnational model to the annual historic data, summing the monthly subnational historic data to the annual level, and then predict the monthly national ACM for 2020–2021 using the $p_t$ fit on the historic annual data.

## 5.2 Annual Data Model

We have annual national ACM counts for Viet Nam, China and Grenada. For these countries we estimate the monthly counts using a multinomial model. This model is derived from the overdispersed Poisson model (4) that is used for countries with no pandemic data. Conditioning on an annual total leads to a multinomial model for the monthly ACM within-year counts with apportionment probabilities $E_{c,t}\theta_{c,t}/\sum_{t'=1}^{12} E_{c,t'}\theta_{c,t'}$, where $\theta_{c,t}$ is given by the log-linear covariate model (5). To obtain counts for these countries, we sample expected numbers $E_{c,t}$ and rates $\theta_{c,t}$from the posterior and subsequently sample multinomial counts with these probabilities.

# 6 Methods for deriving the age- and sex-pattern of excess deaths

Beyond determining the levels of excess mortality attributable to COVID-19, we intend to disaggregate these deaths by age and sex. Unfortunately, for most countries, the mortality data available for the years 2020 and 2021 do not include information on age and sex. This information is crucial for understanding how COVID-19 has affected different age and sex groups, how it compares to other causes of death, and how it has impacted life expectancy. In particular, these are important for generating WHO Global Health estimates for the current year and projections within the World Population Prospects.

To generate estimates of excess mortality by age and sex, we compare the expected age sex distribution for the years 2020 and 2021 with the reported data from places where such information is available. For a specific location, the reported (or observed) death numbers for a particular age-group $x$, sex $s$, country $c$ in year $y$, are represented as

$$Y_{c,y,s,x}$$

These numbers are assumed to be the result of both direct and indirect effects of COVID-19, i.e. the deaths directly attributable to it and the indirect knock-on effects on health systems and society.

To determine the expected number of deaths in the absence of COVID-19 (i.e. how many deaths would have occurred in a no-COVID-19 scenario), we use historical data to forecast the number of deaths, which is represented as

$$E_{c,y,s,x}$$

Excess deaths by age and sex, represented by $\delta_{c,y,s,x}$, can be calculated as the difference between the reported and the expected death numbers :

$$\delta_{c,y,s,x} = Y_{c,y,s,x} - E_{c,y,s,x}$$

As with the approach taken for deaths over all ages and for both sexes combined, the goal is to identify standard patterns of excess mortality by age and sex in places with reported data and then use these patterns to estimate excess mortality by age and sex in countries where such information is not available. Simultaneously, we aim to propagate the uncertainty in the overall excess death numbers for the years 2020 and 2021 (predicted using the statistical models for overall mortality) to the predicted age- and sex- patterns. More details on the steps required to accomplish this can be found in the following sections.

## 6.1 Countries with observed age-sex data for years 2020 and 2021

We consider country- and sex-specific deaths for the year 2020 aggregated to broad age-bands $x \in \{0, 1-4, 5-9, \ldots, 95+\}$. Of interest is the location- and year-$t$-specific death-rate in age interval $x$,represented by $m_{x,t}$. This rate is calculated using the available counts and population numbers from the World Population Prospects (WPP). However, data for this level of granularity is only available for a subset of countries for the years 2020 and 2021. Of these, we have excluded the countries that have experienced conflict, have very small populations, incomplete death data, and/or erratic/implausible age-patterns. The countries with age-sex patterns that we have incorporated into our model framework are listed in Table 2 below.

Table 2: *Countries with age and sex data used in model*

| | WHO Region | ISO3 | Population | Deaths 2020 | Deaths 2021 |
|---|---|---|---|---|---|
| 1 | AFR | CPV | 582664 | 2959 | |
| 2 | AFR | MUS | 1297836 | 11060 | 13274 |
| 3 | AFR | ZAF | 58801925 | 583954 | 711556 |
| 4 | AMR | ARG | 45036031 | 377234 | |
| 5 | AMR | BOL | 11936172 | 74863 | 84301 |
| 6 | AMR | BRA | 213196297 | 1556824 | 1855622 |
| 7 | AMR | CAN | 37888711 | 307702 | 311956 |
| 8 | AMR | CHL | 19300310 | 126169 | 137426 |
| 9 | AMR | COL | 50930668 | 301432 | 363078 |
| 10 | AMR | CRI | 5123118 | 26209 | 31081 |
| 11 | AMR | CUB | 11300699 | 112449 | 167645 |
| 12 | AMR | DOM | 10999673 | 45861 | 47799 |
| 13 | AMR | ECU | 17588595 | 116397 | 105321 |
| 14 | AMR | GTM | 17362729 | 96066 | 118093 |
| 15 | AMR | MEX | 125998305 | 1071243 | 1102123 |
| 16 | AMR | NIC | 6755901 | 33681 | |
| 17 | AMR | PAN | 4294407 | 23876 | |
| 18 | AMR | PER | 33304754 | 240919 | 264086 |
| 19 | AMR | PRY | 6618699 | 35263 | 52711 |
| 20 | AMR | URY | 3429084 | 32638 | 41168 |
| 21 | AMR | USA | 335941996 | 3390165 | 3471759 |
| 22 | EMR | BHR | 1477490 | 3488 | |
| 23 | EMR | EGY | 107465141 | 666200 | |
| 24 | EMR | IRN | 87290197 | 505080 | 531068 |
| 25 | EMR | IRQ | 42556991 | 152901 | |
| 26 | EMR | KWT | 4360455 | 10569 | |

| | | | | | |
|---|---|---|---|---|---|
| 27 | EMR | OMN | 4543407 | 10589 | 12649 |
| 28 | EMR | TUN | 12161729 | 75058 | |
| 29 | EUR | ALB | 2866849 | 27605 | 30580 |
| 30 | EUR | AUT | 8907775 | 89879 | 89864 |
| 31 | EUR | AZE | 10284962 | 75647 | 76878 |
| 32 | EUR | BEL | 11561717 | 126748 | 112303 |
| 33 | EUR | BGR | 6979184 | 124703 | 149040 |
| 34 | EUR | BIH | 3318423 | 42803 | |
| 35 | EUR | CHE | 8638611 | 75994 | 71010 |
| 36 | EUR | CZE | 10530961 | 129327 | 139876 |
| 37 | EUR | DEU | 83328993 | 985106 | 1023653 |
| 38 | EUR | DNK | 5825640 | 54621 | 57214 |
| 39 | EUR | ESP | 47363802 | 492287 | 449041 |
| 40 | EUR | EST | 1329450 | 15706 | 18355 |
| 41 | EUR | FIN | 5529472 | 55435 | 57703 |
| 42 | EUR | FRA | 64480059 | 654540 | 643748 |
| 43 | EUR | GBR | 67059477 | 687182 | 667318 |
| 44 | EUR | GEO | 3765911 | 50537 | 59906 |
| 45 | EUR | GRC | 10512232 | 131033 | 143792 |
| 46 | EUR | HRV | 4096876 | 56992 | 62873 |
| 47 | EUR | HUN | 9750582 | 140997 | 155573 |
| 48 | EUR | IRL | 4946128 | 32856 | |
| 49 | EUR | ISR | 8757489 | 48797 | 50756 |
| 50 | EUR | ITA | 59500577 | 746146 | 709035 |
| 51 | EUR | KAZ | 18979246 | 162613 | 183357 |
| 52 | EUR | KGZ | 6424882 | 39977 | |
| 53 | EUR | LTU | 2820277 | 43516 | 47758 |
| 54 | EUR | LVA | 1897052 | 28825 | 34587 |
| 55 | EUR | MDA | 3084860 | 40716 | 45386 |
| 56 | EUR | NLD | 17434561 | 168609 | 170756 |
| 57 | EUR | NOR | 5379846 | 40598 | 42003 |
| 58 | EUR | POL | 38428368 | 478138 | 519877 |
| 59 | EUR | PRT | 10298190 | 123685 | 125188 |
| 60 | EUR | ROU | 19442042 | 296880 | 335530 |
| 61 | EUR | RUS | 145617325 | 2138586 | 2441594 |
| 62 | EUR | SRB | 7358011 | 116850 | 135901 |
| 63 | EUR | SVK | 5456689 | 59089 | 73461 |
| 64 | EUR | SVN | 2117651 | 23960 | 23764 |
| 65 | EUR | SWE | 10368965 | 95412 | 89019 |
| 66 | EUR | UKR | 43909669 | 616835 | 714263 |
| 67 | EUR | UZB | 33526665 | 175637 | 174540 |
| 68 | SEAR | THA | 71475663 | 501438 | 563650 |
| 69 | WPR | AUS | 25670047 | 162639 | 172012 |

| | | | | | |
|---|---|---|---|---|---|
| 70 | WPR | BRN | 441750 | 1752 | |
| 71 | WPR | JPN | 125244761 | 1384544 | 1452289 |
| 72 | WPR | KOR | 51844686 | 305049 | 317506 |
| 73 | WPR | MNG | 3294344 | 15922 | 19931 |
| 74 | WPR | MYS | 33199999 | 166507 | |
| 75 | WPR | NZL | 5061135 | 32738 | 34908 |

## 6.2 Grouping countries to generalize age-sex patterns

We need to determine how to group the countries with available data on age and sex so that we can use that information to extrapolate the age-sex impact of COVID-19 on locations without data. One possible way would be to group countries geographically, using regional identification such as WHO regions. However, this approach presents two challenges:

1. Some regions are not well represented in the available observed data, for example, there are only a few countries from the AFR and SEAR regions.
2. Countries that are geographically close can have vastly different impacts of the pandemic, including by age and sex. For instance, Finland and Russia share borders, but the excess death rates for these countries are significantly different.

Instead of relying on the natural geography to group the data, we allow the data to drive the grouping. We apply the K -means clustering approach (Likas *et al.*, 2003), which is commonly used to partition a dataset into K groups automatically. The K -means method characterizes the data by using K centers of clusters. Each observation $x_i$ is assigned to a given cluster such as the sum of squares distance of the observation to their assigned cluster centers ($m_k$) is minimised. We defined the total within-cluster variation as follows;

$$J_K = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - m_k)^2$$

, where $x_i$ is a data point belonging to the cluster $C_k$ and $m_k$ is defined as the mean value of the points assigned to the cluster $C_k$

$m_k = \sum_{i \in C_k} \frac{x_i}{n_k}$ ($n_k$ being the number of points in $C_k$).

There are numerous options of features that could be utilized to construct the data matrix $X = (x_1, \ldots, x_{n)}$, which is used to inform the clusters. Four features are considered in total (one of which relates to the predicted excess mortality attributable to COVID-19):

1. For each country, the human development index for year 2019:

$$f_1 = HDI$$

2. For each country, the mean age at death in 2019:

$$f_2 = \frac{\sum_x x \times Y_{x,2019}}{\sum_x Y_{x,2019}}$$

, where $Y_{x,2019}$ refers to the age-specific all causes deaths in 2019.

3. For each country, the crude excess mortality rate estimated for year $t$:

$$f_{3,t} = \frac{\delta_t}{\sum_x N_{x,t}}$$

, where $\delta_t$ represents the predicted overall excess deaths and $N_{x,t}$ corresponds to the age-specific population numbers.

4. For each country, the excess deaths relative to the total all causes number of deaths for year t (as a percentage):

$$f_{4,t} = \frac{\delta_t}{Y_t}$$

, where $Y_t$ represents the estimated or reported total number of deaths.

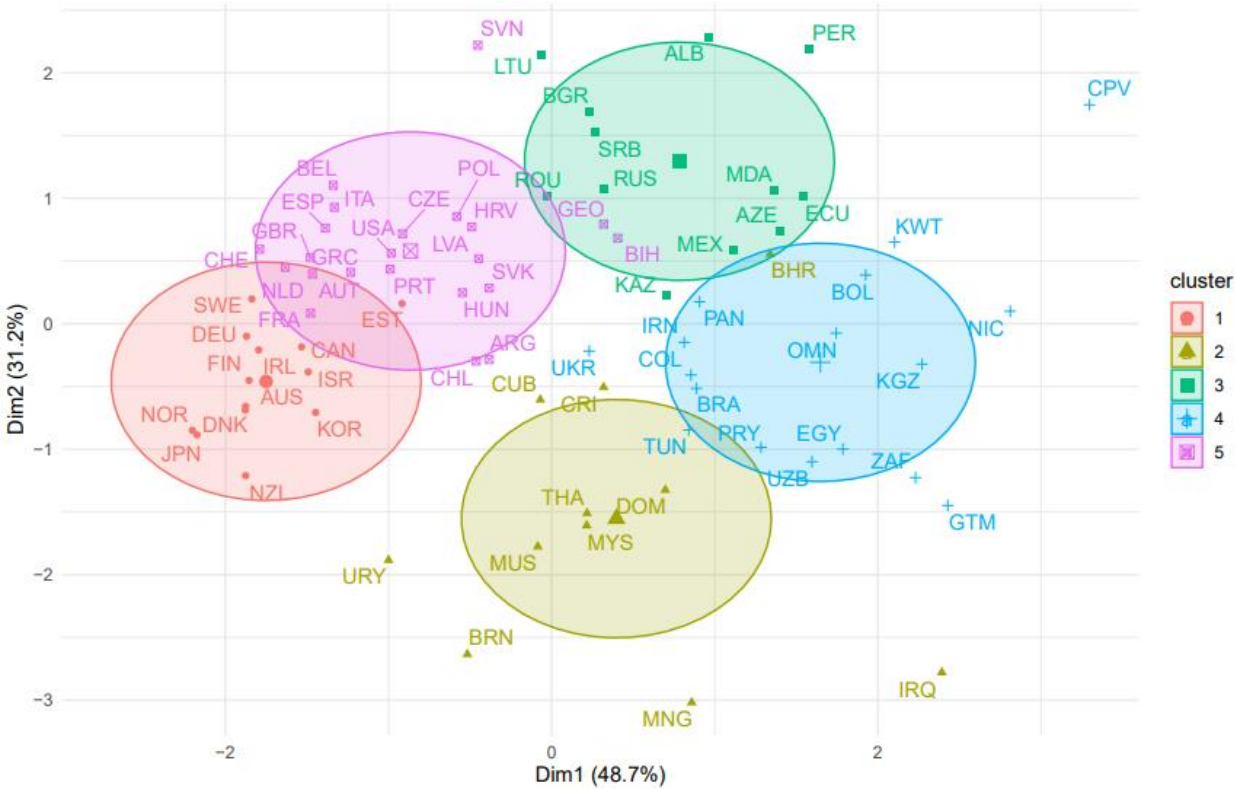
These are derived for each WHO member state and normalized to derive the X matrix:

$$X_{(194,4)} = \begin{bmatrix} f_{1,1} & f_{2,1} & f_{3,1} & f_{4,1} \\ \vdots & \vdots & \vdots & \vdots \\ f_{j,1} & f_{j,2} & f_{j,3} & f_{j,4} \\ \vdots & \vdots & \vdots & \vdots \\ f_{194,1} & f_{194,2} & f_{194,3} & f_{194,4} \end{bmatrix}$$

The K-means approach is used for the subset of countries with available observed sex- and age-specific data for the year 2020. The number of clusters $K$ is chosen in order that it maximises the variation between clusters and minimises the variation within clusters. For each year, five clusters are generated. The resulting cluster compositions are presented in Figure 5 and tables 3 and 4 list the respective cluster constituents.

Figure 5: *K-means cluster allocation by ISO3 for years 2020 and 2021*

K−means clusters: (hdi, mean age at death, excess rate and percent), 2020




K−means clusters: (hdi, mean age at death, excess rate and percent), 2021

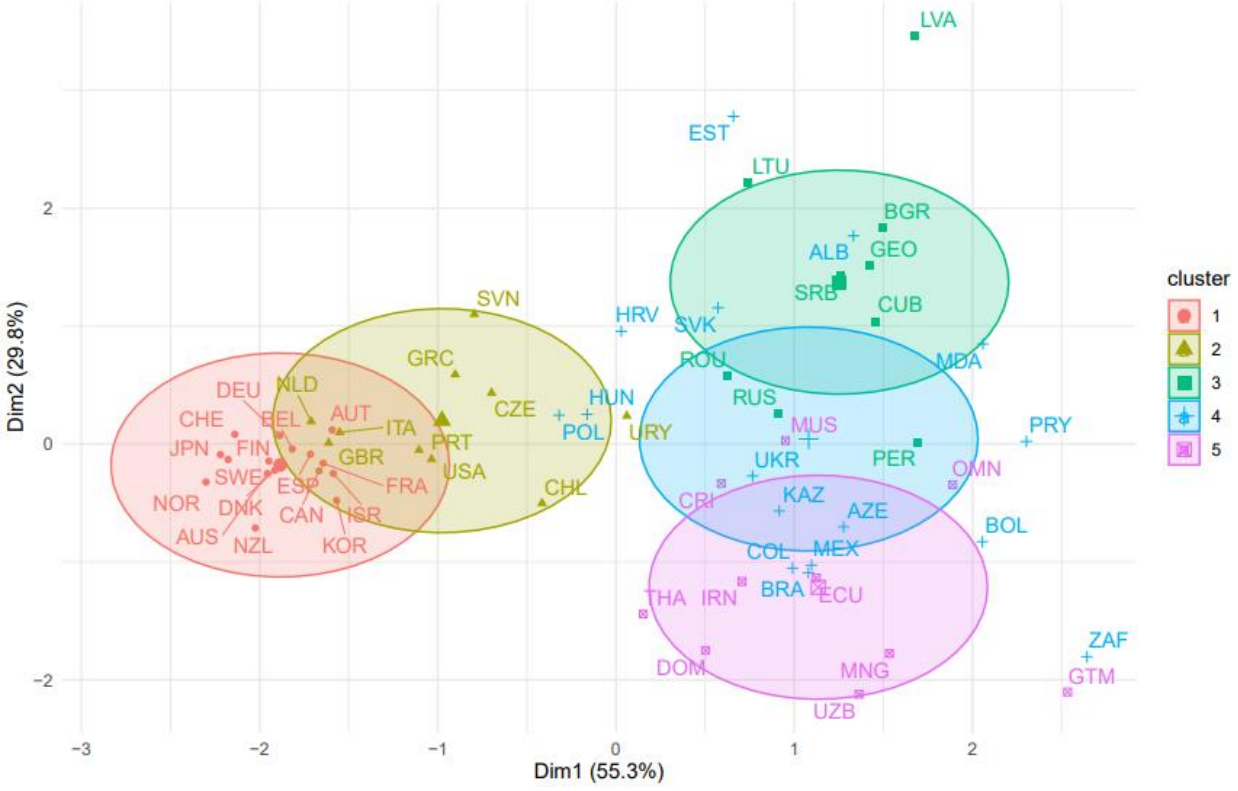


Table 3: *K-means cluster allocation for 2020*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| AUS | BHR | ALB | BOL | ARG |
| CAN | BRN | AZE | BRA | AUT |
| DEU | CRI | BGR | COL | BEL |
| DNK | CUB | ECU | CPV | BIH |
| EST | DOM | KAZ | EGY | CHE |
| FIN | IRQ | LTU | GTM | CHL |
| IRL | MNG | MDA | IRN | CZE |
| ISR | MUS | MEX | KGZ | ESP |
| JPN | MYS | PER | KWT | FRA |
| KOR | THA | ROU | NIC | GBR |
| NOR | URY | RUS | OMN | GEO |
| NZL | | SRB | PAN | GRC |
| SWE | | | PRY | HRV |
| | | | TUN | HUN |
| | | | UKR | ITA |
| | | | UZB | LVA |
| | | | ZAF | NLD |
| | | | | POL |
| | | | | PRT |
| | | | | SVK |
| | | | | SVN |
| | | | | USA |

Table 4: *K-means cluster allocation for 2021*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| AUS | CHL | BGR | ALB | CRI |
| AUT | CZE | CUB | AZE | DOM |
| BEL | GBR | GEO | BOL | ECU |
| CAN | GRC | LTU | BRA | GTM |
| CHE | ITA | LVA | COL | IRN |
| DEU | NLD | PER | EST | MNG |
| DNK | PRT | ROU | HRV | MUS |
| ESP | SVN | RUS | HUN | OMN |
| FIN | URY | SRB | KAZ | THA |
| FRA | USA | | MDA | UZB |
| ISR | | | MEX | |
| JPN | | | POL | |
| KOR | | | PRY | |
| NOR | | | SVK | |
| NZL | | | UKR | |
| SWE | | | ZAF | |

## 6.3 Extrapolating cluster groupings to countries without observed data

The K -means clustering approach provides country groupings based on multiple factors, including the HDI, mean age at death, overall excess mortality rates and proportion of total deaths. Not all countries are included in the clustering process as this was done for a subset to ensure that all countries with data are clustered into optimal bins.

However, we require all countries to be assigned to clusters and this is accomplished by mapping each country to the 5 K -mean clusters using the multivariate Minkowski distance (Singh *et al.*, 2013) between the X matrix values and the cluster averages shown in the Tables 5 and 6 :

Table 5: *K-means cluster average normalised values for year 2020*

| Cluster | HDI | Mean age at death | Crude excess rate | Excess proportion |
|---|---|---|---|---|
| 1 | 1.4727 | 1.1525 | -0.7759 | -0.1116 |
| 2 | 0.5609 | 0.3652 | -0.9604 | -0.1124 |
| 3 | 0.5371 | 0.6141 | 1.1889 | -0.0983 |
| 4 | 0.0438 | 0.1689 | 0.0111 | -0.1083 |
| 5 | 1.1461 | 1.1343 | 0.2317 | -0.1073 |

Table 6: *K-means cluster average normalised values for year 2021*

| Cluster | HDI | Mean age at death | Crude excess rate | Excess proportion |
|---|---|---|---|---|
| 1 | 1.4251 | 1.2418 | -0.5803 | -0.1109 |
| 2 | 1.1869 | 1.1680 | 0.3972 | -0.1068 |
| 3 | 0.6426 | 0.8916 | 3.4557 | -0.0847 |
| 4 | 0.3908 | 0.4680 | 1.6320 | -0.0979 |
| 5 | 0.3160 | 0.3397 | 0.1116 | -0.1019 |

## 6.4 Extrapolating age-sex impact to countries without observed data

In order to assess excess mortality due to COVID-19 in 2020 and 2021, both for countries with observed data and those without, it is necessary to determine the expected number of deaths by age and sex. These expected number of deaths by age and sex are derived using the reported deaths prior to 2020 (where available, otherwise GHE2019 if no data is available) forecasted to 2020/2021 using the age-period Lee-Carter model (AP-LC).

Consider a mortality experience observed at individual ages $x$ and calendar years $t$, giving rise to a total of $(k \times n)$ observations, so that we can estimate the central mortality rate $(m_{xt})$ and the corresponding force of mortality( $\mu_{xt}$) by

$$(\hat{\mu}_{xt} =)\hat{m}_{xt} = \frac{y_{xt}}{e_{xt}},$$

where $y_{xt}$ and $e_{xt}$ represent the number of deaths and the matching central exposure for any given subgroup, respectively. In addition, for each combination of age $x$ and period $t$, we define the cohort year $z = t - x$ representing the year of birth of each subgroup in the data.

The AP LC model is expressed as

$$\log(m_{xt}) = \alpha_x + \beta_x \kappa_t + \varepsilon_{xt},$$

where the parameters are interpreted as follows:

- $\alpha_x$ represents a constant age-specific pattern of mortality;

- $\kappa_t$measures the trend in mortality over time;
- $\beta_x$ measures the age-specific deviations of mortality change from the overall trend;
- $\varepsilon_{xt}$reflects age and time effects not captured in the model and are Gaussian distributed $N(0, \sigma^2)$.

The expected number of deaths by age and sex are adjusted to align with the expected death numbers obtained from the overall longitudinal mortality model. This approach ensures that the expected number of deaths are consistent and reduces any potential bias that may arise from relying solely on age and sex forecasts for the two-year period.

For the countries with observed data listed in the table above ( excluding those with low death counts or zero inflated counts) and for each age and sex-group, we look at the scale $r_{s,x}$ representing the age- and sex-specific percentage change in all-causes mortality when comparing the observed to the expected i.e. the P-Score:

$$r_{s,x} = 100 \times \left( \frac{Y_{s,x} - E_{s,x}}{E_{s,x}} \right)$$

,where $Y_{s,x}$ and $E_{s,x}$ refer to the observed and expected deaths, respectively. This quantity contrasts the observed against the expected to capture the age- and sex-specific changes for 2020/2021. The K -means clusters serve two purposes in extrapolating data for countries without observed data. Firstly, they are used to summarise these log mortality scalars into cluster specific distributions. Secondly, they are used to derive country-specific estimates of predicted deaths by age and sex, taking into account the clusters in which the country is classified.

For each cluster $k$ (and by extension, each country $j$ in the cluster $k$), we generate sex-specific distribution for the $r_{s,x}$scalars based on the observed data. The empirical bootstrap distribution is generated by first smoothing the observed series by age for each country in the cluster and then repeatedly sampling from the smooth series. The range of possible values by age is assumed to be a Gaussian approximate with distribution

$$\hat{r}_{s,x}^{k} \sim N(\bar{r}_{s,x}^{k}, \sigma_{s,x}^{k})$$

,where $\bar{r}_{s,x}^{k}$and $\sigma_{s,x}^{k}$ are the sex $s$ and age $x$ specific mean and standard deviations for cluster $k$ derived using the smoothed draws of the observed data.
Following recommendations from the United Nations Interagency Group for Child Mortality Estimation (UNIGME), we do not extrapolate any protection or otherwise to children and young adults under age 25. Figures 6 to 10 are of the observed and smoothed series and are filtered by cluster.

Figure 6: *Smoothed ratio by age and sex for clusters 1 to 2 in 2020*

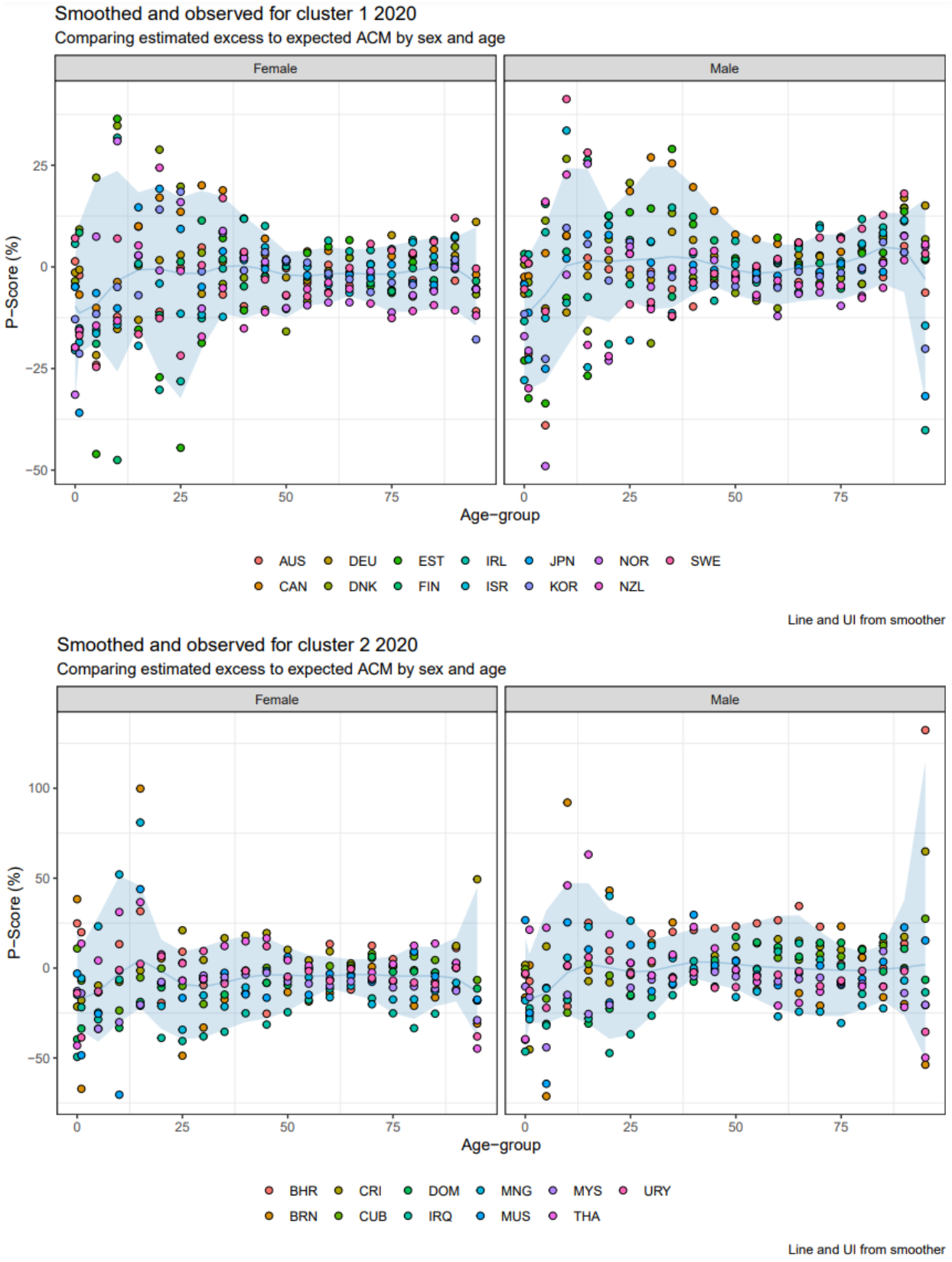
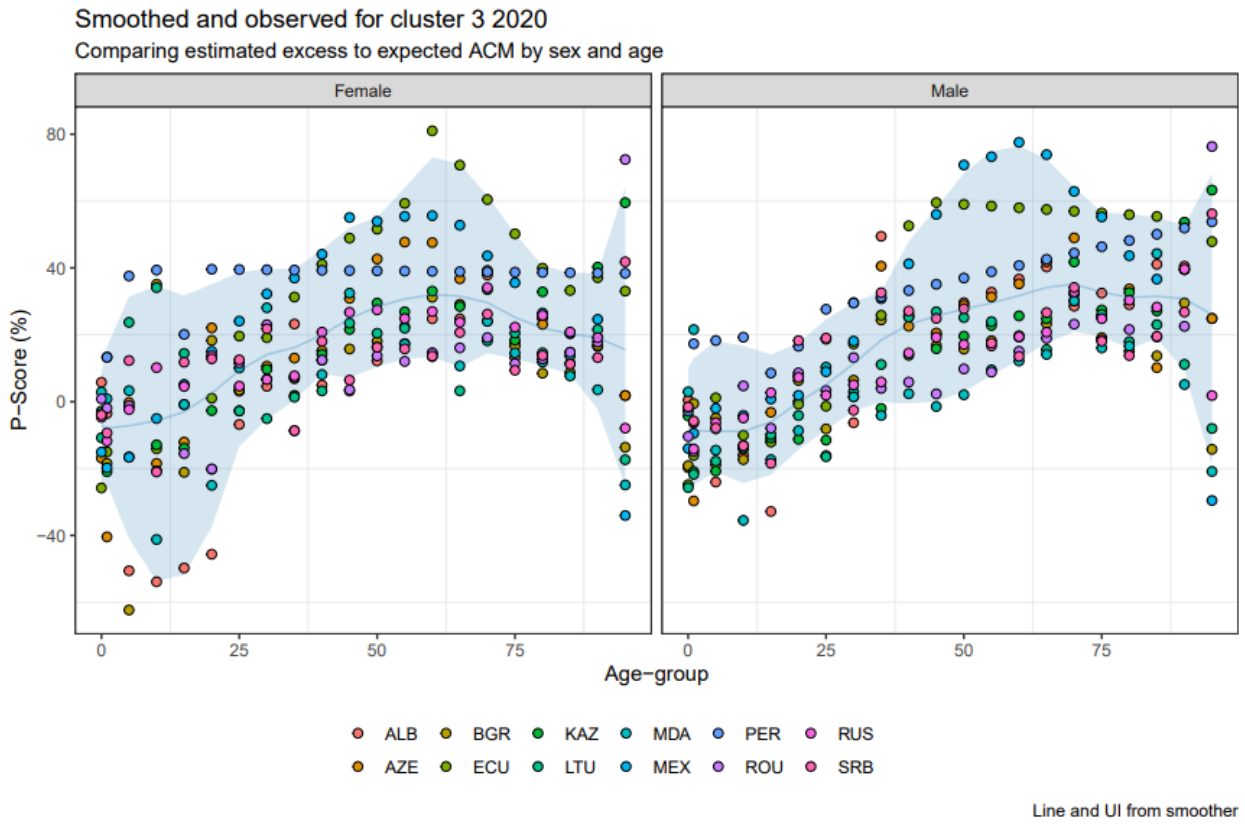


Figure 7: *Smoothed ratio by age and sex for clusters 3 to 4 in 2020*



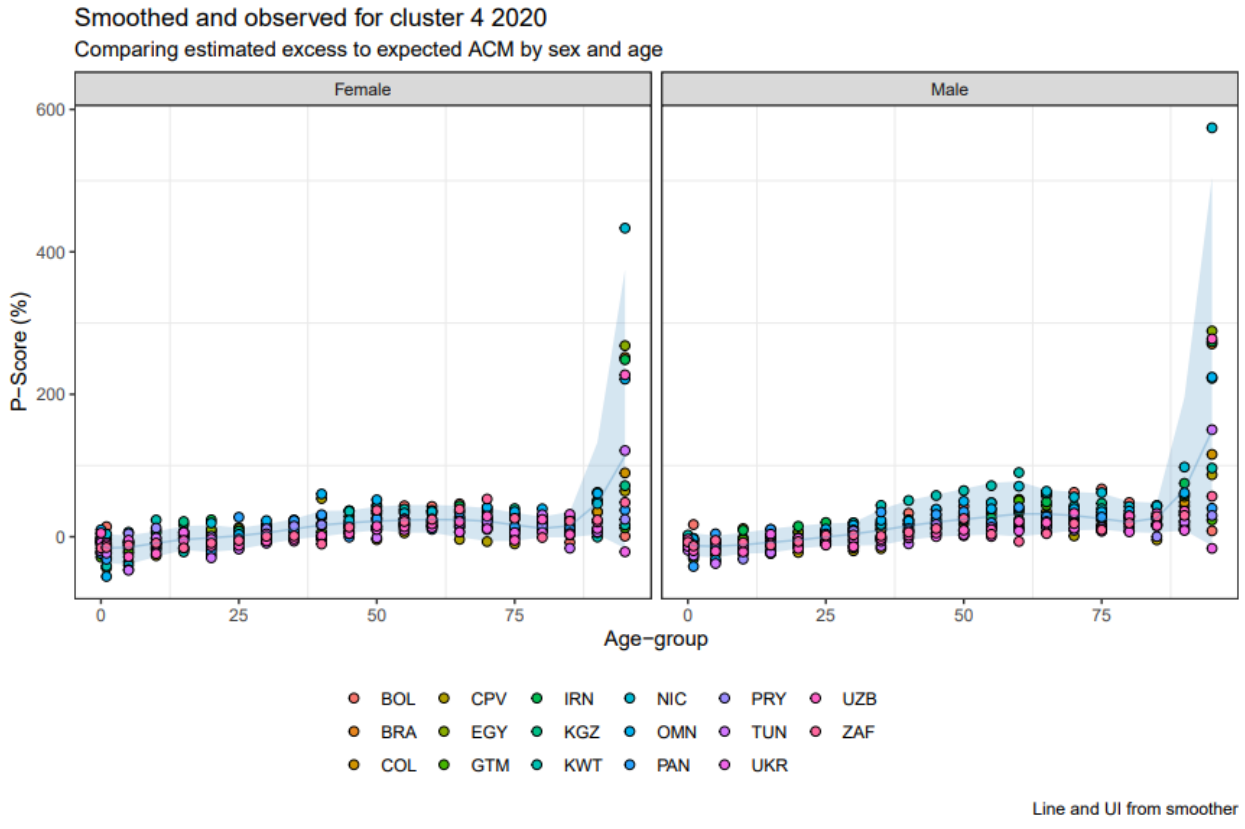



Figure 8: *Smoothed ratio by age and sex for cluster 5 2020 and cluster 1 in 2021*

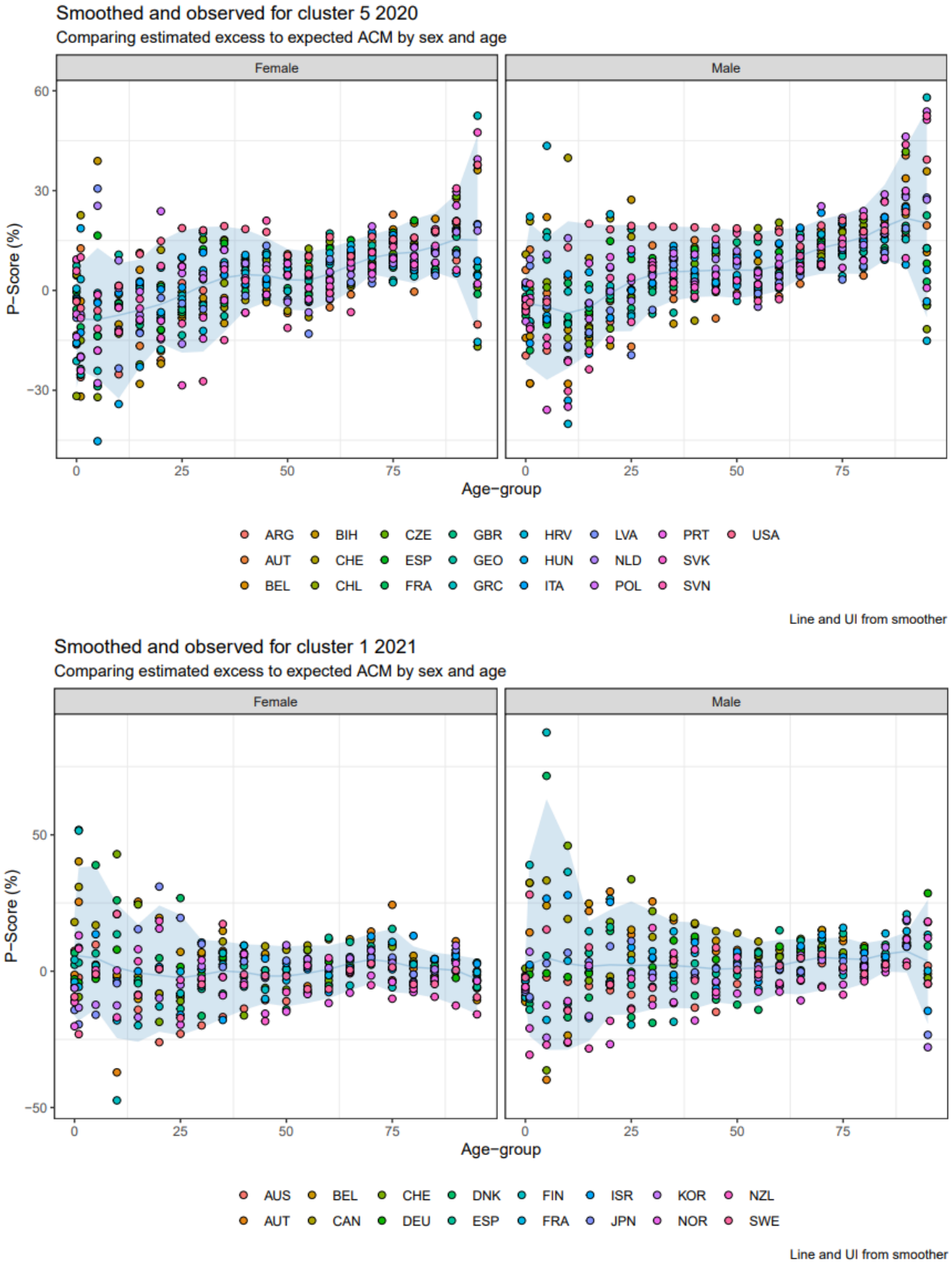


Figure 9: *Smoothed ratio by age and sex for clusters 2 to 3 in 2021*

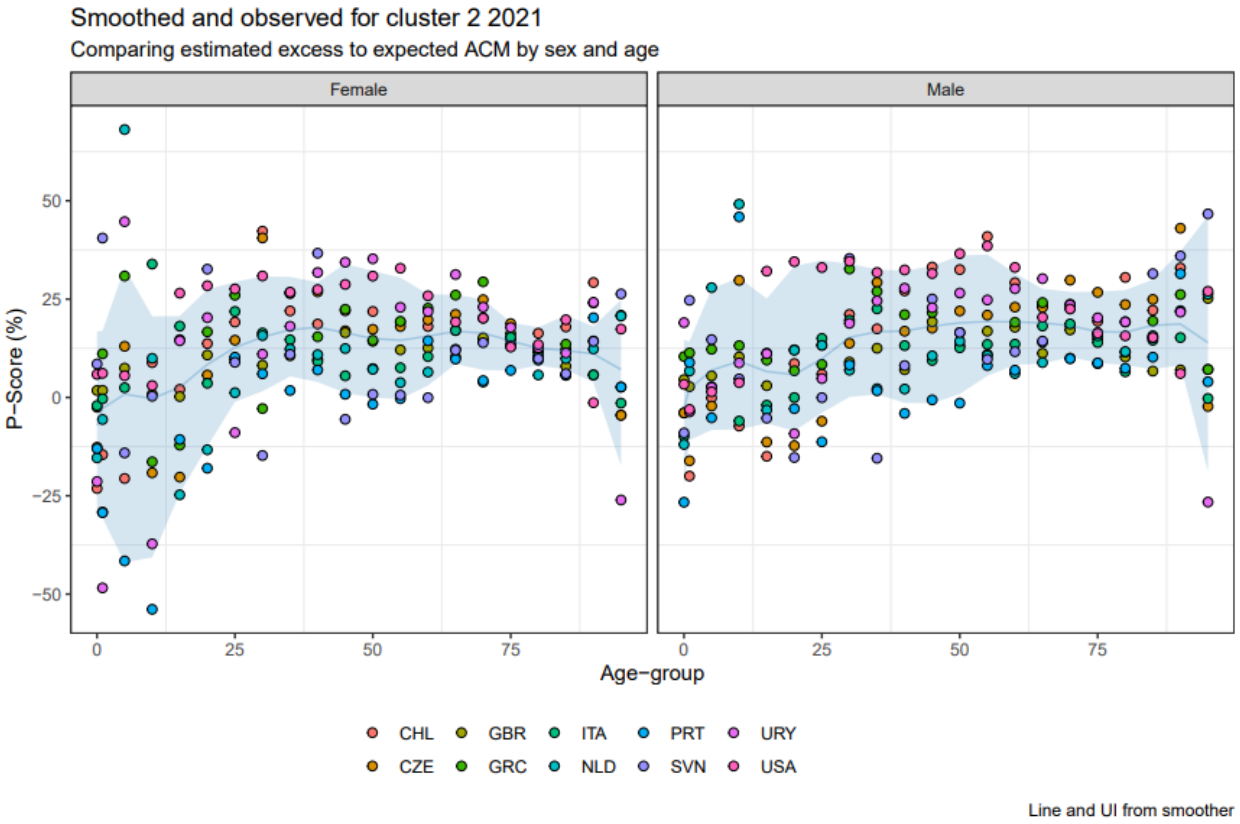
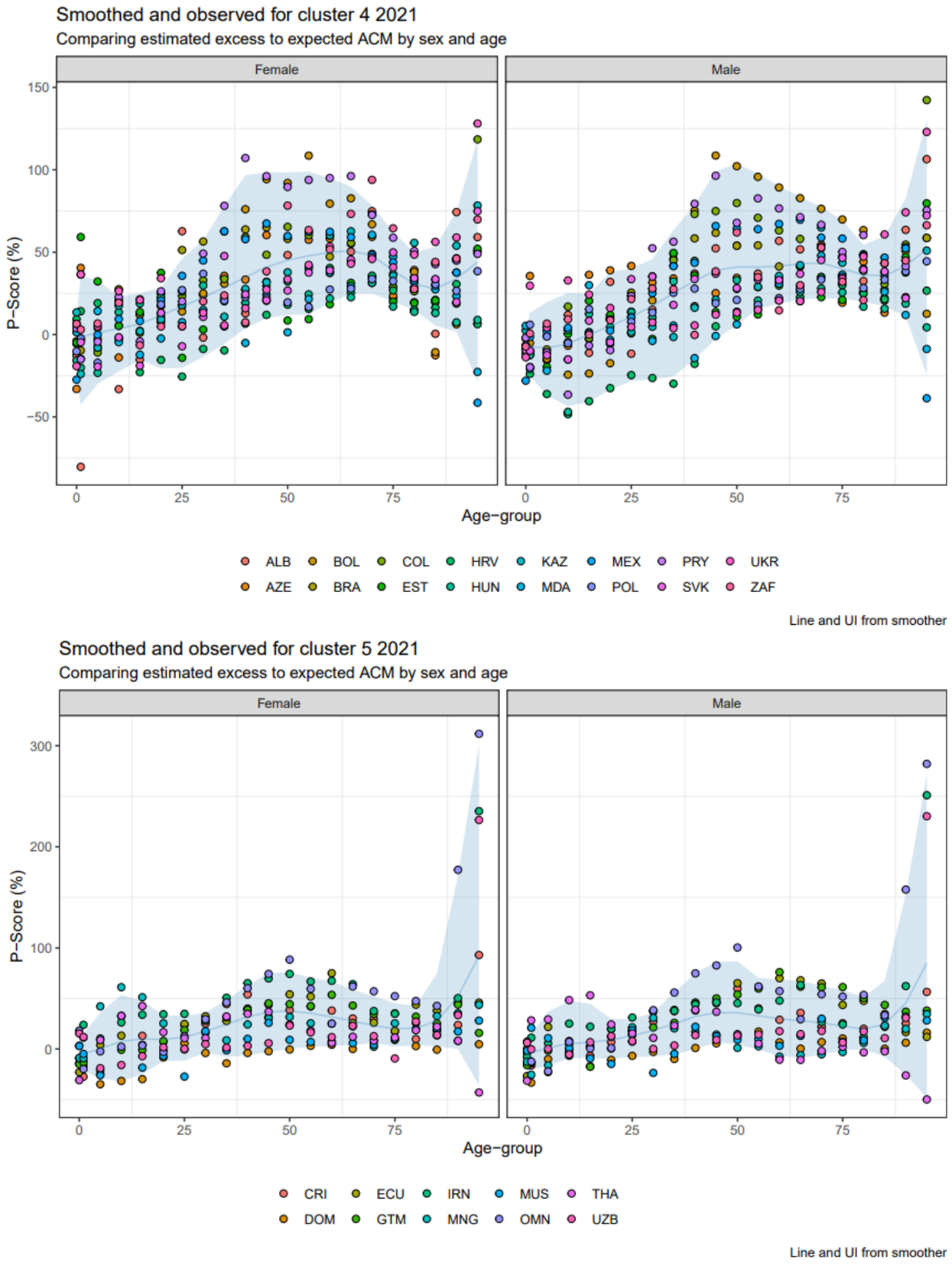



Figure 10: *Smoothed ratio by age and sex for clusters 4 to 5 in 2021*



Finally, the draws of the scalars are combined with samples of the age $x$ and sex $s$ specific expected deaths by country to generate predicted deaths by age and sex:

$\widehat{D}_{s,x} = \widehat{E}_{s,x}(1 + \frac{\hat{r}_{s,x}}{100})$

Each $\widehat{D^{\wedge}}$ vector is rescaled to correspond to a random realization from the Poisson count model $\widehat{Y^{\wedge}}$ i.e.

$$\widetilde{D}_{s,x} = \frac{\widehat{D}_{s,x}}{\sum_s \sum_x \widehat{D}_{s,x}} \times \hat{Y}$$

This process is repeated 1,000 times for each country, drawing unique samples of $\hat{r}_{s,x}$, $\hat{E} \sim N(\bar{E}, \sigma_E)$ and $\hat{Y} \sim N(\bar{Y}, \sigma_{Y)}$ each time. These are used to generate country-specific distributions (and uncertainty intervals) of deaths by age and sex. The uncertainty shown is the propagation of the uncertainty from the K -means cluster smoothed draws and the Poisson count model draws but should not be interpreted as being parametric or containing a hypothetical ”true” value. Instead, it depicts a range of plausible values depending on the total predicted deaths’ distribution, the expected deaths and the cluster assigned to the country.

# 7 Removal of deaths from other shocks

When calculating excess deaths, it is important to distinguish between deaths that are directly related to COVID-19 (directly or indirectly) and deaths that may be caused by other shocks, such as deaths from conflicts and natural disasters. Therefore, it is common practice to remove deaths from other shocks from the total number of excess deaths to obtain a more accurate estimate of the impact of COVID-19 on mortality.

The United Nations Population Division consolidated global datasets that described the number of deaths due to (1) conflicts, including wars, mass killings (including genocides), battle deaths, etc. and (2) natural disasters, such as floods, cyclones, epidemics, earthquakes, famines, droughts and tsunamis. From the consolidated data, estimates of the number of deaths from each crisis event by country and year were computed, with associated uncertainty bounds. Various analytical methods were applied to reconcile overlaps across different data sources, gaps and discontinuities due to changes in definitions or territorial boundaries of countries, non-annual reference periods, and sources that provided a broad range of crisis-related deaths rather than a point estimate, among other challenges. More details on the methodology can be found at WPP2022_Methodology.pdf. The data used are available at: WPP2022_F02_METADATA.XLSX.

# 8 Links to statistical code, input and output data

R-syntax and data for generating estimates and replicating parts of the analyses can be found at https://github.com/WHOexcessc19/Codebase2022Update.

# 9 Results

The results obtained through the utilization of this method can be accessed at: https://www.who.int/data/sets/global-excess-deaths-associated-with-covid-19-modelled-estimates
The results published in May 2022 have been discussed in the following paper : Msemburi, W., Karlinsky, A., Knutson, V. et al (2023). The WHO estimates of excess mortality associated with the COVID-19 pandemic. Nature 613, 130–137.

# 10 Acknowledgments

The World Health Organization would like to acknowledge and extend its appreciation to Professor Jon Wakefield and his team comprised of Victoria Knutson and Dr. Serge Aleshin-Guendel from University of Washington for their exceptional work in both developing and authoring the methodology presented in sections 3, 4, and 5 of this document. More details about those sections can be found in the following scientific paper :

Knutson, V., Aleshin-Guendel, S., Karlinsky, A., Msemburi, W., and Wakefield, J. (2023). “Estimating global and country-specific excess mortality during the COVID-19 pandemic”. Annals of Applied Statistics, 17, 1353-1374.

# References


Adair, T. and Lopez, A. D. (2018). Estimating the completeness of death registration: An empirical method. *PLOS ONE*, 13, e0197047.

Anand, S. and Sen, A. (1994). Human Development Index: Methodology and Measurement.

Baker, S. G. (1994). The multinomial-Poisson transformation. *Journal of the Royal Statistical Society: Series D*, 43, 495–504.

Checchi, F. and Roberts, L. (2005). HPN network paper 52: Interpreting and using mortality data in humanitarian emergencies: a primer for non-epidemiologists. Technical report, Overseas Development Institute.

GBD (2020). Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, 396, 1204–1222.

Hale, T., Angrist, N., Kira, B., Petherick, A., Phillips, T., and Webster, S. (2020). Variation in government responses to COVID-19. Technical report, University of Oxford.

Karlinsky, A. (2021). International completeness of death registration 2015-2019. *medRxiv*.

Karlinsky, A. (2022). Estimating national excess mortality from subnational data: application to Argentina. *Revista Panamericana de Salud Publica*.

Karlinsky, A. and Kobak, D. (2021). Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. *eLife*, 10, e69336.

Kelsall, J. E., Zeger, S. L., and Samet, J. M. (1999). Frequency domain log-linear models; air pollution and mortality. *Journal of the Royal Statistical Society: Series C*, 48, 331–344.

Knutson, V., Aleshin-Guendel, S., Karlinsky, A., Msemburi, W., and Wakefield, J. (2023). "Estimating global and country-specific excess mortality during the COVID-19 pandemic. *Annals of Applied Statistics,* 17, 1353-1374.

Kung, S., Doppen, M., Black, M., Hills, T., and Kearns, N. (2020). Reduced mortality in New Zealand during the COVID-19 pandemic. *The Lancet*.

Leon, D. A., Shkolnikov, V. M., Smeeth, L., Magnus, P., Pechholdová, M., and Jarvis, C. I. (2020). Covid-19: a need for real-time monitoring of weekly excess deaths. *The Lancet*, 395, e81.

Likas, A., Vlassis, N., and Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, 36(2), 451–461.

Link, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial N. *Ecology*, 94, 2173–2179.

Mikkelsen, L., Phillips, D. E., AbouZahr, C., Setel, P. W., Savigny, D., Lozano, R., and Lopez, A. D. (2015). A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. *The Lancet*, 386, 1395–1406.

Msemburi, W., Karlinsky, A., Knutson, V. et al (2023). The WHO estimates of excess mortality associated with the COVID-19 pandemic. Nature 613, 130–137.

Németh, L., Jdanov, D. A., and Shkolnikov, V. M. (2021). An open-sourced, web-based application to analyze weekly excess mortality based on the Short-term Mortality Fluctuations data series. *PloS ONE*, 16, e0246663.

Parks, R. M., Bennett, J. E., Foreman, K. J., Toumi, R., and Ezzati, M. (2018). National and regional seasonal dynamics of all-cause and cause-specific mortality in the usa from 1980 to 2016. *Elife*, 7, e35500.

Riffe, T. and Acosta, E. (2021). Data resource profile: COVerAGE-DB: a global demographic database of COVID-19 cases and deaths. *International Journal of Epidemiology*, 50, 390–390f.

Rivera, R., Rosenbaum, J. E., and Quispe, W. (2020). Excess mortality in the United States during the first three months of the covid-19 pandemic. *Epidemiology and Infection*, 148.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Application*. Chapman and Hall/CRC Press, Boca Raton.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B*, 71, 319–392.

Simpson, D., Rue, H., Riebler, A., Martins, T., and Sørbye, S. (2017). Penalising model component complexity: A principled, practical approach to constructing priors (with discussion). *Statistical Science*, 32, 1–28.

Singh, A., Yadav, A., and Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, 67(10).

UNSD (2021). Demographic Yearbook 2021. Technical report, United Nations Statistics Division.

United Nations, Department of Economic and Social Affairs, Population Division (2022). World Population Prospects 2022, Online Edition.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R, Second Edition*. CRC Press.