Methodology for estimating global excess deaths associated with COVID-19

Last updated on 16 December 2021
CONTENTS

1 Introduction 3
2 Methods for estimating all-cause deaths 4
  2.1 Expected deaths in countries with historical time-series by week or month 4
  2.2 Expected deaths in countries with historical time-series by year only 5
  2.3 Predicted deaths in countries without reported data for January 2020 to June 2021 6
3 Methods for deriving sex- and age-pattern 7
  3.1 Grouping countries to generalize sex-age patterns 8
  3.2 Extrapolating cluster groupings to countries without observed data 9
  3.3 Extrapolating 2020 sex-age impact to countries without observed data 10
  3.4 Propagating the uncertainty from the all-cause model 11
1 INTRODUCTION

The World Health Organisation has been tracking the impact of COVID-19 as the pandemic has evolved over time. Aggregate case and death numbers are being reported to the WHO and the data have been made publicly available\(^1\). These data do not always provide a complete picture of the health burden attributable to COVID-19 nor of how many lives have been lost due to the pandemic, and there are a number of reasons for this. Some deaths that are attributable to COVID-19 have not been certified as such because tests had not been conducted prior to death. There have also been variations in the death certification rules countries have applied in the presence of comorbidities and COVID-19. The impact of the pandemic is far reaching. Beyond the deaths directly attributable to it are those that can be linked to the conditions that have prevailed since the pandemic began and have led to some health systems being overwhelmed or some patients avoiding healthcare. In light of the challenges posed by using reported COVID-19 data, excess mortality is considered a more objective and comparable measure\(^2\). The WHO defines\(^3\) excess mortality as "the mortality above what would be expected based on the non-crisis mortality rate in the population of interest".

To derive estimates of the excess mortality attributable to COVID-19, the location \(c\) all-causes death numbers for year 2020/21 and week or month \(t\), can be represented by

\[
D^r_{c,t}
\]

and are assumed to be a result of the direct effects of COVID-19 (deaths attributable to it) and the indirect knock-on effects on health systems and society. The hypothetical or "counterfactual" no-COVID-19 scenario uses the expected death numbers

\[
D^e_{c,t}
\]

which are forecasted using the historical location and time specific deaths (prior to the pandemic). Excess deaths, represented by \(\omega_{c,t}\), can thus be defined as the difference:

\[
\omega_{c,t} = D^r_{c,t} - D^e_{c,t}
\]

The exercise of determining excess deaths is non-trivial due to the data required to do so. Currently, excess mortality cannot be derived for all countries when using standard methods because of gaps in the underlying data. Routine mortality data is usually received by the WHO a year or more after the year of death or following an even longer lag. In addition, differential reporting capacity and variable data quality across countries has resulted in many nations lacking the systems to provide good quality routine data even in the past. Correspondingly,

\(^1\) Source: https://covid19.who.int/
\(^2\) https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)30933-8/fulltext
\(^3\) https://www.who.int/hac/about/definitions/en/
these countries lack the capacity and data required to monitor all-cause mortality during this unprecedented pandemic. Consequently, a number of countries are 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. Apart from data gaps in all-cause death numbers for years 2020 and 2021, many are missing crucial attributes such as sex and age. These dimensions are critical for standardised comparisons across countries that account for underlying age-structure and for estimating life tables. In addition, a large number of countries do not have historical data that are disaggregated by periods shorter than a year. This document provides an overview of the methods that have been used to generate excess mortality by the WHO for the period January 2020 to June 2021. Additional methodological detail, input data, all results and the R-syntax used to generate them will follow.

2 METHODS FOR ESTIMATING ALL-CAUSE DEATHS

2.1 Expected deaths in countries with historical time-series by week or month

For each location we assume that $D_t$, which represents the death count for week or month $t$, is a random variable that follows a negative-binominal distribution with mean parameter $\mu_t$, i.e.,

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

where $\phi$ is an overdispersion parameter such that $\text{Var}(D_t) = \mu_t + \mu_t^2 / \phi$ and with

$$\log(\mu_t) = \alpha + \beta_t + \gamma(t)$$

where $\alpha$ is an intercept, $\beta_t$ tracks the annual trend, and $\gamma(t)$ is a smooth function of time $t$ which accounts for within year seasonal variation. $\gamma(t)$ is fit using a penalised cyclic cubic regression spline\(^4,5\).

The model is fit to the location data for the period 2015 to 2019 and used to predict the expected deaths $\hat{\mu}_t$ for all $t$ in years 2015 to 2021. Approximate simulations from the Bayesian posterior density are used to generate 95% credibility intervals of the expected deaths. Distributions for the estimated excess deaths $\omega$ are generated using draws of the expected i.e.,

$$\omega_t = D_t - \hat{\mu}_t$$

such that cumulative excess deaths are just the aggregate $\sum_t \omega_t$.

---

2.2 Expected deaths in countries with historical time-series by year only

For countries with only annual historical data, the goal is to determine excess deaths by month $t$ for $t \in 1, \ldots, 18$ which correspond to the entirety of year 2020 and the first half of 2021. As stated previously, the calculation requires two specific inputs i.e., the expected deaths for time $t$ ($D^e_{c,t}$) and the actual deaths for time $t$ ($D_{r,t}$). For just under half of all WHO member states (tier 1 data availability), the actual deaths are captured through routine mortality surveillance systems. For these countries, their historical data can be forecasted using standard time-series methods to derive expected deaths by month for the years 2020/2021, which can in turn be used to calculate excess deaths directly. However, for the remaining countries without the data (tier 2 data availability), statistical models are applied to generate both the expected and the actual. These models are described in detail in the following sections.

Expected deaths by month and country in the tier 2 countries are generated over two steps:

(i) Firstly, country-specific annual series of all-cause mortality aggregated across all-ages and for both sexes combined are obtained from the Global Health Estimates (GHE2019). These deaths, $D_y$, for years $y \in 2000, \ldots, 2019$ are modelled using a second order random walk (RW2). The RW2 model has the density

$$
\pi(D) \propto \exp \left( -\frac{1}{2} \sum_{i=3}^{n} (D_y - 2D_{y-1} + D_{y-2})^2 \right)
$$

where the term $D_{y+1} - 2D_y + D_{y-1}$ can be interpreted as an estimate of the second order derivative of a continuous time function $D(z)$ at $z = y$ using values of $D(z)$ at $z = y-1$, $y$, and $y+1$. Using this model, we forecast the annual deaths for years 2020 and 2021 which can be represented by $D^{e*}$.

(ii) Next, the predicted annual series is dis-aggregated to month using the inverse relationship between the seasonal patterns and monthly average death count as observed in the countries with reported data. We use a smooth series of the monthly temperatures since 2015 as proxy for season and normalise the temperatures across countries by removing the mean and scaling by the standard deviation. For each location we assume that $D^e_t$, which represents the expected death count in month $t$, is a random variable that follows a negative-binomial distribution conditional on the annual expected deaths $D^{e*}$ and a mean parameter $\mu_t$ i.e.,

$$
D^e_t|\mu_t \sim \text{NegBin}(\frac{D^{e*}}{n}, \mu_t, \phi)
$$

where $\phi$ is an overdispersion parameter and with

$$
\log(\mu_t) = \alpha + W + \gamma(t) + \beta_t x_t
$$

where $\alpha$ is an intercept, $W$ is a dummy variable for WHO region, $\gamma(t)$ is a smooth function on time $t$ across the 18 months, and $\beta_t$ is a coefficient on the normalised temperature $x_t$.
which is allowed to be time-varying via a random walk of order 2 (RW2) prior. Here we take \( n = 12 \) to calculate the average number of deaths per month conditional on \( D_e^\ast \).

### 2.3 Predicted deaths in countries without reported data for January 2020 to June 2021

The second component of the excess deaths is the number of actual deaths for 2020/2021 by month \( t \). These are predicted according to the following algorithm. For a select country \( c \), we let:

- \( N_c \) be the population size in 2020/2021.
- \( Y_{ct} \) be the total deaths in month \( t \) (observed in a subset of countries and assumed to be complete with good coverage), \( t = 1, \ldots, T \) with \( T = 18 \) (numbered from beginning of 2020).
- \( D_{ct}^e \) be (modeled) expected deaths in 2020, in the absence of COVID-19 (generated for all countries).

We apply an expected offset model of the form:

\[
Y_{ct} | \theta_{ct} \sim \text{Poisson}(D_{ct}^e \theta_{ct})
\]

which has \( E[Y_{ct} | \theta_{ct}] = D_{ct}^e \theta_{ct} \), so that \( \theta_{ct} > 0 \) is a relative rate parameter. If this parameter is \(<1/>1 \) then the death rate in country \( c \) and in month \( t \) is less than that/greater than that expected, based on historical data.

We model the relative rate as:

\[
\log \theta_{ct} = \beta_0 + \sum_{b=1}^{B} \beta_{b}^v X_{ctb} + \sum_{g=1}^{G} \beta_{g}^c Z_{cg} + \epsilon_{ct}
\]

where:

- \( \beta_0 \) is global intercept, (note this differs from the models with different intercepts for countries with different income levels e.g., high/not high income).
- \( \epsilon_{ct} \sim iid N(0, \sigma_{\epsilon}^2) \) is a random effect which allows for excess-Poisson variation around the covariate model.
- We have \( B \) time-varying covariates,
  - \text{sqrt(COVID-19)}: The square root of the reported COVID-19.\(^6\)
  - \text{positivity rate}: Percentage of COVID-19 tests that are positive, 0 to 100.\(^7\)

\(^6\) Source: https://covid19.who.int/
\(^7\) Source: https://ourworldindata.org/placeholder/positive-rate-daily-smoothed
– containment and health index: which combines 'lockdown' restrictions and closures 0 to 100.8

We allow the associations with these variables, $\beta_{bl}$ to be time-varying, via a random walk of order 2 (RW2) prior. These parameters do not include a sum-to-zero constraint, since we want the centered around the overall (across all $T$ months) association.

• The $G$ time-invariant covariate (a dummy variables for whether or not the country is a high income economy according to the 2021 World Bank classification) with constant association parameters $\beta_{g}^{C}$.

3 METHODS FOR DERIVING SEX- AND AGE-PATTERN

Beyond determining the levels of excess mortality attributable to COVID-19, we disaggregate these deaths by age and sex. For most countries, the sex and age attributes are not identified in the mortality data that are available for years 2020 and 2021. These are necessary inputs when one begins to look at sex and age differentials, impact relative to other causes as well as life tables, and the impact on life expectancy.

To generate estimates of excess mortality by age and sex we consider the expected sex and age profile for the year 2020 and assess the difference in the expected and the observed in the places with reported data. For a specific location, the observed/reported death numbers for an age-group $x$ and sex $s$ in year 2020, can be represented by

$$D_{s,x}^{r}$$

and are assumed to be a result of the direct effects of COVID-19 (deaths attributable to it) and the indirect knock-on effects on health systems and society. The hypothetical or "counterfactual" no-COVID-19 scenario uses the expected death numbers

$$D_{s,x}^{e}$$

which are forecasted using historical data. Excess deaths by sex and age, represented by $\omega_{s,x}$, can thus be defined as the difference:

$$\omega_{s,x} = D_{s,x}^{r} - D_{s,x}^{e}$$

As with the approach taken for deaths over all ages and for both sexes combined, the goal is to determine standard patterns for sex and age for the places with reported data and then generalize them to the other countries without. Simultaneously, we aim to propagate the

8 Source: https://www.bsg.ox.ac.uk/research/research-projects/covid-19-government-response-tracker
uncertainty in the overall death numbers for the year 2020 (predicted using the statistical models for overall mortality) to the predicted sex- and age-patterns.

We consider country- and sex-specific deaths for the year 2020 aggregated to 5-year age-bands \( x \in \{0 - 4, 5 - 9, \ldots, 85 + \} \). Of interest is the location- and year-specific death-rate in age interval \([x, x + n]\), represented by \( n m_{x,t} \) which is calculated using the counts and the population numbers according to WPP2019. Only a subset of countries has observed data to estimate these quantities at this level of granularity for 2020.

### 3.1 Grouping countries to generalize sex-age patterns

We consider how to group the countries with data in order to extrapolate any of the estimated 2020 impacts to the locations without. A natural grouping would be geographically using some regional identification e.g., the WHO region. However, this poses two dilemmas:

1. Not all regions are represented adequately in the observed data e.g., there are no countries from SEARO region and there is only one country from AFRO.

2. Even within close geographic proximity, the scale of the impact of the pandemic overall and thus potentially by sex and age, varies e.g., Finland shares borders with Russia but the reported excess rates for the countries is on significantly different orders if magnitude.

Instead of using the natural geography to group the data, we apply the \( K \)-means clustering approach. \( K \)-means is a method commonly used to automatically partition a data set into \( K \) groups. The \( K \)-means method uses \( K \) centers of clusters, to characterize the data. These centers are determined by minimizing the sum of squared errors,

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

where \((x_1, \ldots, x_n) = X\) is the data matrix and \(m_k = \frac{\sum_{i \in C_k} x_i}{n_k}\) is the centroid of cluster \(C_k\) and \(n_k\) is the number of points in \(C_k\).

There are many options of features that could be used to create the data matrix \(X\) which is informing the clusters. Using the age-specific all-cause deaths in 2019, \(D_x\), the mean predicted total deaths in 2020 \(\bar{Y}\), age-specific population in 2019 and 2020, \(N_x\) and \(N_x^*\) respectively, and the predicted overall excess deaths for 2020, \(\gamma\), we consider four features in total (two of which are measures of the historical underlying age pattern and the other two are measures related to the predicted excess mortality attributable to COVID-19):

(i) For each country, mean age at death in 2019:
\[ a_d = \frac{\sum x \times D_x}{\sum D_x} \]

(ii) For each country, mean age of the population in 2019:
\[ a_p = \frac{\sum x \times N_x}{\sum N_x} \]

(iii) For each country, the crude excess mortality rate for 2020:
\[ \gamma_r = \frac{\gamma}{\sum N_x} \]

(iv) For each country, the excess mortality as a proportion of mean predicted total deaths for 2020:
\[ \gamma_p = \frac{\gamma}{\bar{\gamma}} \]

These are derived for each of the 194 member states and normalized to derive the X matrix:

\[ X_{(194,4)} = \begin{bmatrix} j & 1 & 2 & 3 & 4 \\ 1 & a_{d,1} & a_{p,1} & \gamma_{r,1} & \gamma_{p,1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ j & a_{d,j} & a_{p,j} & \gamma_{r,j} & \gamma_{p,j} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 194 & a_{d,194} & a_{p,194} & \gamma_{r,194} & \gamma_{p,194} \end{bmatrix} \]

### 3.2 Extrapolating cluster groupings to countries without observed data

The K-means clusters provide groupings for the data based on mean ages at death and of population, as well as the overall excess mortality rates and proportion of total deaths. Not all countries are included in the original clustering 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 K-mean clusters using the multivariate minkowski\(^{10}\) distance between the X matrix values and the cluster averages. Essentially, a country \( j \) is assigned to the cluster for which the minkowski distance between these cluster averages and the vector \( X_{j,\cdot} \) are at a minimum.

---

3.3 Extrapolating 2020 sex-age impact to countries without observed data

To capture the Covid-19 excess mortality impact for the year 2020 for the countries with observed data as well as those without, requires expected death numbers by sex and age. For the expected deaths we simply rescale the GHE2019 sex and age distribution to the 2020 expected death numbers that are derived for the overall longitudinal mortality model. This enforces consistency in the expected numbers and also minimizes any potential bias that can be introduced in the single year forecast that would otherwise be necessary to derive the expected deaths for 2020. Acknowledging the uncertainty inherent in the GHE2019 estimates and that this is a very short forecast, we assume that changes in the sex- and age-pattern between observed 2019 and expected 2020 are minimal after adjusting for the level through the aggregate projection.

For the countries with data and listed in the table above (we exclude the countries with low death numbers and zero inflated counts), for each sex and by age-group, we look at the scale $r_{s,x}$:

$$r_{s,x} = \log \left( \frac{n_{ms,x}^o}{n_{ms,x}^e} \right)$$

where the superscripts $o$ and $e$ are for the observed and expected mortality rates, respectively. This quantity contrasts the observed log mortality rate against the expected to capture the sex- and age-specific changes for 2020. The $K$-means clusters are used for two aspects of the extrapolation to locations without data. Firstly, to summarise these log mortality scalars into cluster specific distributions. And secondly, to derive country-specific estimates of predicted deaths by sex and age, conditional on the clusters the country lies in.

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

$$r_{s,x}^k \sim N \left( \mu_{s,x}^k, \sigma_{s,x}^k \right)$$

where $\mu_{s,x}^k$ 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.
3.4 Propagating the uncertainty from the all-cause model

Finally, the draws of the scalars are combined with the sex \( s \) and age \( x \) specific expected death rates by country and the sex-age-specific population numbers to generate predicted deaths by age and sex:

\[
\hat{D}_{s,x} = n \times m_{s,x}^e \times \exp(\hat{\epsilon}_{s,x}) \times N_{s,x}^*
\]

Each \( \hat{D} \) vector is rescaled to correspond to a random realization from the Poisson count model \( \hat{Y} \) i.e.

\[
\hat{D}_{s,x} = \frac{\hat{D}_{s,x}}{\sum_s \sum_x \hat{D}_{s,x}} \times \hat{Y}
\]

This process is repeated 1,000 times for each country, drawing unique samples of \( \hat{\epsilon}_{s,x} \) 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 sex and age. 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. Rather it shows a range of plausible values conditional on the distribution of the total predicted deaths, the expected deaths and the cluster identified for the country.