Updated methods for

Estimating the early death toll of COVID-19 in the United States (May 28, 2020)

Daniel M. Weinberger PhD, Jenny Chen, Ted Cohen, MD DPH, Forrest W. Crawford PhD, Farzad Mostashari MD, Don Olson MPH, Virginia E Pitzer ScD, Nicholas G Reich PhD, Marcus Russi BS, Lone Simonsen PhD, Anne Watkins BS, Cecile Viboud PhD

Summary of major changes from previous versions

1) The data from NCHS, rather than CDC fluview, are being used. NCHS assigns the deaths based on state of occurrence, while CDC fluview assigns the deaths by on state of residence
2) The variable for influenza activity is now de-seasonalized using the Serfling method, and we use increases of influenza above the seasonal baseline as a covariate in the model for excess deaths. Flu activity dropped to historically low levels in Spring 2020, and this adjustment prevents the flu variable from artificially pulling the baseline down (and inflating the estimates of excess deaths).
3) We include an adjustment for reporting delays, which are estimated based on provisional deaths reported weekly since March 2020. This adjustment is applied to the baseline, so more recent weeks that are likely to have incomplete data have a baseline (expected reported deaths) that is adjusted downwards.

Methods Overview

Data

Data on deaths due to pneumonia and influenza and coronavirus ("PIC", ICD-10 codes in the range of U07.1 or J09-J18) and all causes were obtained from the National Center for Health Statistics' (NCHS) mortality surveillance system.[9] Data were stratified by state and week. While excess PIC mortality captures the direct burden of the novel coronavirus, excess all-cause mortality estimates the full impact of the pandemic and could encompass deaths directly caused by the virus but that were not attributed to the virus as well as deaths that were not directly related to the virus. For instance, if people avoid receiving healthcare for chronic conditions, there could be increases in certain categories of deaths. Data on all-cause deaths in previous years were obtained from https://data.cdc.gov/resource/pp7x-dyj2 and

https://data.cdc.gov/resource/muzy-jte6. Data on all-cause deaths and PIC deaths since January 26, 2020 were obtained from https://data.cdc.gov/resource/r8kw-7aab. The NCHS data are based on the state where the death occurred rather than the state of residence.

Historical data on the proportion of deaths due to pneumonia and influenza (P&I) in previous years were obtained from the CDC's weekly P&I deaths reports (https://gis.cdc.gov/grasp/fluview/mortality.html) via the cdcfluview package in R, and these were used to determine the number of pneumonia and influenza deaths in the baseline period. All data were accessed May 22, 2020.

Connecticut and North Carolina were missing mortality data for recent months and were therefore excluded from the analyses and from the baseline numbers.

We also compiled data on COVID-19-related morbidity to gauge the timing and intensity of the pandemic in different locations. We used CDC data on influenza-like illness (ILI) [10], a longstanding indicator of morbidity from acute respiratory pathogens, including SARS-CoV-2. We also obtained information on influenza virus circulation to adjust baseline estimates[11] (see Appendix for details).

To compare our excess mortality estimates with official COVID-19 tallies, we compiled weekly numbers of reported deaths due to COVID-19 in each state from two sources, including the Covid Tracking Project[12] and NCHS[13]. State-specific testing information was obtained from the Covid Tracking Project[12]. The data from the Covid Tracking Project are mostly based on the date of death report, which might lag behind the actual date of death.

Excess mortality and morbidity analysis

To calculate the number of COVID-19-related excess deaths, we subtracted the expected number of deaths in each week from the observed number of deaths for the period March 1, 2020 to May 9, 2020. Each of the 48 states (except North Carolina and Connecticut) and the District of Columbia were analyzed individually. We fit Poisson regression models to the weekly state-level death counts from January 5, 2015 to January 25, 2020 (see Supplement for details). The baseline was then projected forward until May 9, 2020; excess mortality was defined as the observed mortality minus the baseline. The baseline was adjusted for seasonality, year-to-year

baseline variation, influenza epidemics, and reporting delays. The PIC mortality model used all-cause deaths as a denominator. Poisson 95% prediction intervals were estimated by sampling from the uncertainty distributions for the estimated model parameters.[14] Pennsylvania was not highlighted in the figures, despite having a large number of excess deaths, because the data were incomplete during March 2020. To obtain national-level estimates, the observed count and baseline counts were summed for each week and compared.

Evaluating reporting delays

We aestimated the reporting delays using a modified version of the NobBS package in R and incorporated that as an adjustment in the main analysis[16] (Supplementary methods). The completeness of the data varied markedly between states (**Figure S1**).

Code and data availability

The analyses were run using R v3.6.1. All analysis scripts and archives of the data are available from https://github.com/weinbergerlab/excess_pi_covid

More detailed methods

Datasets:

Details on the choice of mortality indicators

We used multiple cause of death data to extract deaths with PIC causes (pneumonia, influenza , coronavirus) listed anywhere on the death certificate. Excess mortality from pneumonia and influenza has been used in the US to monitor the severity of influenza since the 1918 pandemic. Here we concentrate on PIC mortality rather than pneumonia alone, or P&I alone, to be more comprehensive. Deaths coded as 'influenza' or 'coronavirus' do not necessarily require laboratory confirmation of infection, and there is overlap of symptoms between influenza and COVID-19. The PIC grouping includes individuals with a cause of death listed as COVID-19 (either with or without a P&I code) as well as people who did not have COVID-19 listed but did have a cause of death of pneumonia or influenza. PIC codes could be present alone or in combination.

Reported COVID-19 deaths

The number of COVID-19 deaths reported to NCHS were used for most states. NCHS suppresses data when there are 1-9 counts for a particular week/state—this was an issue for a few

of the smaller states. In these instances, the data from the Covid tracking project for that week and state were substituted in. In more recent weeks, the data from the Covid tracking project were higher than the official tallies in many states, likely due to shorter reporting delays. Therefore, this substitution would have the effect of shrinking the gap between excess deaths and reported COVID-19 deaths. Because this is only an issue in states with small counts, it does not meaningfully change the overall estimates.

Baseline data on pneumonia and influenza deaths

For the pre-pandemic period, there were two datasets that were combined to get the number of pneumonia and influenza deaths. NCHS provides data on all-cause deaths and pneumonia and influenza deaths by week and state for previous years based on *underlying cause*. CDC fluview produces estimates of the number and proportion of deaths due to pneumonia and influenza by state and week that is based on *multiple-cause* of death coding (pneumonia/influenza code anywhere on the death certificate). But the CDC fluview data are based on state of residence, rather than state of occurrence. To get an estimate of pneumonia and influenza deaths by state of occurrence and by multiple cause of death that would be comparable with the pneumonia/influenza/coronavirus definition produced for 2020, we multiplied the proportion of deaths due to P&I from CDC FluView with the number of all-cause deaths reported by NCHS.

Data on morbidity and circulation of other respiratory pathogens:

Weekly state-level ILI data were obtained from the CDC's ILINet system[10], which aggregates data from a network of outpatient providers. To adjust for activity of non-SARS-CoV-2 respiratory pathogens, we used state-level data on laboratory-confirmed influenza activity from the CDC's National Respiratory and Enteric Virus Surveillance System (NREVSS)[11]. This dataset captures the number of tests performed for influenza and the number that were positive by week and state. The ILI data provide the percent of visits to participating outpatient providers that were for ILI. ILINet and NREVSS data are available with a 1-week lag.

The ILI, NREVSS, and P&I mortality datasets were accessed through the CDC's FluView portal using the cdcfluview package in R. Data from NCHS, ILINet and NREVSS were obtained for the weeks ending January 5, 2013 through May 9, 2020.

Comparison with changes in influenza-like illness

To get a measurement of COVID-19 epidemic intensity in the outpatient setting, the same model was fit to ILI data. Time series for excess ILI were compared with times series for excess all-cause deaths.

Adjustments to the influenza confirmed cases time series

In our main analyses, it was desirable to adjust for influenza activity when estimating the seasonal baseline. There are two reasons for this. First, failure to adjust for influenza epidemics that were still ongoing in the Spring would lead to over-attributing excess deaths to influenza. Second, influenza epidemics in previous years could bias the seasonal baseline upwards. However, starting in March 2020 the number of influenza cases and hospitalizations in the US dropped to historically low levels. Because of this drop, including influenza activity in the model could bias the baseline downward and lead to an over-estimate of excess deaths. To avoid this issue, we quantified increases in influenza activity above a seasonal baseline. Briefly, the percent of specimens positive for influenza were log transformed, we used a "Serfling" regression approach to set a seasonal baseline. Data for flu season (December-February) were set to missing, and we fit a linear regression to the remaining data, with harmonics with periods of 52 and 26 weeks. We then subtracting the fitted harmonic baseline from the observed data, and any negative values were set to 0. This provides a time series of increases of influenza above the seasonal baseline while ignoring recent decline in influenza below the baseline. This time series was used as a covariate in the main analysis.

Statistical model

We fit Poisson regression models to the weekly state-level death counts from January 5, 2015 to January 25, 2020. The baseline was then projected forward until May 9, 2020; excess mortality was estimated as observed minus baseline deaths. Models were fit separately for each state. We adjusted for seasonality, year-to-year baseline variation, and influenza activity in the previous week (details above). The 1-week lag between the influenza data and the mortality data accounts for the delay between the time when the influenza test is performed and death. Influenza data for Florida were not reported, so we used influenza data for the other states in region 4 (southeastern US) instead.

We regress all-cause deaths and PIC deaths in epidemiological year *i* (July-June) and week *t,* using the equation below (shown for all-cause deaths). The PIC model was the same, except that all-cause deaths were used as an offset-term, and there was no adjustment for Proportion Complete.

Let $\text{Total\_Deaths}_{i,t}$ be the number of deaths and let $\text{Flu\_Epidemic}_{i,t-1}$ be the time series of influenza epidemic activity (see above), and $\text{Proportion\_Complete}_{i,t}$ the estimated proportion of deaths that have been reported for that state and week (see next section) We modeled

$$\text{Total\_Deaths}_{i,t} \sim \text{Poisson}(\boldsymbol{\lambda}_{i,t,}, \phi)$$

where

$$\log(\boldsymbol{\lambda}_{i,t}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1*\sin(\Theta_t) + \boldsymbol{\beta}_2*\cos(\Theta_t) + \boldsymbol{\beta}_3*\sin(\Theta_t/2) + \boldsymbol{\beta}_4*\cos(\Theta_t/2) + \\ + \boldsymbol{\beta}_6*\log(\text{Flu\_Epidemic}_{i,t-1}) + \gamma_i + \alpha_i*\log(\text{Flu\_Epidemic}_{i,t-1}) + \log(\text{Proportion\_Complete}_{i,t})$$

and

$$\Theta_t = 2*\boldsymbol{\pi}*t/52.1775$$

To compute prediction intervals, we used the following procedure. Once the regression coefficients were estimated, we extracted the estimated asymptotic covariance matrix for the parameters and constructed a multivariate normal distribution approximating the sampling

distribution, centered at the estimated parameter values. We drew 100 samples from this parameter distribution, and for each sample computed the resulting mean value $\boldsymbol{\lambda}_{i,t}$, and then drew 100 samples from the Poisson distribution with this mean. This resulted in 10,000 samples from an empirical predictive distribution of $\text{Total\_Deaths}_{i,t}$. Empirical 95% prediction intervals were computed by taking the 2.5th and 97.5th percentiles of this resulting distribution. In sensitivity analyses, we evaluated a model in which Flu_Epidemic was excluded altogether.

Reporting delays analysis

We used an empirical and model-based approach to estimate the reporting delays.

Empirical approach: First, to determine whether reporting delays changed between 2019 and 2020 outside of the period of intense pandemic activity, we compared the provisional death reports filed in week 11 in 2019 and 2020 and found that they did not diverge (**Figure S1A**). This demonstrates that the reporting delays did not change in 2020. We then compared the provisional data reported for weeks 10-19 in week 20 of 2020 in the CDC FluView report with the data reported for the same period in the week 20 of 2019. The area between these two curves is an empirical estimate of excess deaths but is not adjusted for year-year variations in deaths. In January and February 2020, there were an average of 829 more deaths in 2020 compared to 2019, representing increases not related to COVID-19. We summed the area between the 2020 and 2019 curves for weeks 10-19 and subtracted 829*(10 weeks). This provides an empirical estimate of excess deaths, adjusted for reporting delays and year-over-year increases.

Model-based approach: We used the state-level FluView reports for weeks 11-20 in 2020 to construct reporting triangles for each state. Archived versions of the FluView data were not publicly available for earlier weeks. We used the nowcasting with Bayesian smoothing method described by McGough et al (*PLOS Comp Biol 2020)*, using a negative binomial model. The only modification was that the reporting triangle was not complete because we did not fully observe the data. The earliest week of data we considered was 2 weeks after death. Any deaths reported in the first 2 weeks were combined into a single category. For dates of death where the first N weeks were not observed the weekly reporting probabilities for weeks 2:N were combined to estimate the reporting fraction of the observed count. The modified JAGS code is here: https://github.com/weinbergerlab/excess_pi_covid/blob/master/functions/jags_negbin4.R. The proportion reported by each week after death was estimated from the model. The proportion reported each week was combined to get a cumulative estimate of the proportion of deaths that were reported by a particular week. The median of 10,000 draws from the posterior distribution for each delay week was estimated. This estimate of proportion complete was used as a denominator in the main analysis. For earlier weeks with complete data, this value is 1 or close to 1. For more recent weeks, this value is smaller (indicating incomplete data), and the baseline of expected reported deaths is adjusted down accordingly.

The smoothed nowcast estimates of deaths from the NobBS model were not directly used in our analyses. Sensitivity analyses using these nowcasted estimates in place of the observed counts found national estimates of excess deaths that were approximately 4% higher than the estimates presented here.