Estimating Unreported Deaths Associated with COVID-19

Benjamin Stear, M.S.
Department of Biomedical and Health Informatics
The Children's Hospital of Philadelphia
3401 Civic Center Blvd.,
Philadelphia, PA 19104-4399

Kyle M. Hernandez, Ph.D.
Center for Translational Data Science
and
Department of Medicine
University of Chicago,
Chicago, IL.

Vidya Manian, Ph.D.
Department of Electrical and Computer Engineering
& Bioengineering
University of Puerto Rico
PO Box 9000
Mayaguez, PR 00681-9000

Deanne Taylor, Ph.D.
Department of Biomedical and Health Informatics
The Children's Hospital of Philadelphia
and
Department of Pediatrics,
University of Pennsylvania Perelman Medical School
3401 Civic Center Blvd.,
Philadelphia, PA 19104-4399

Catharine A. Conley, Ph.D.
Space Science and Astrobiology Division,
NASA Ames Research Ctr.
Moffett Field, CA 94035
202-288-8718
cassie.conley@nasa.gov

Classification

Biological Sciences; Medical Sciences

Keywords

COVID-19, Epidemiology, Modeling, Estimation

Author Contributions

Dr. Conley had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Conley, Stear, Hernandez.

Acquisition, analysis, or interpretation of data: Conley, Stear, Hernandez, Manian.

Drafting of the manuscript: Conley.

Critical revision of the manuscript for important intellectual content: Hernandez, Manian, Taylor.

Statistical analysis: Conley, Stear.

Supervision: Conley.

Significance Statement

Estimating deaths from natural causes using the percentage of natural cause deaths from COVID-19 reported to the CDC and COVID-19 deaths counted by public tracking sites suggests that up to 200,000 deaths from natural causes between 22 April and 4 August, 2020, 20% of the total recorded as of 26 August, have not yet been reported to the CDC.

Acknowledgments

We thank the founders and members of the COVID-19 International Research Team, Dr.s Afshin Beheshti, Todd Treagan, and Krista Ternus, for facilitating this collaboration; and particularly members of the COV-IRT Modeling Subgroup, particularly Dr.s Daniel Guimarães Tiezzi, Sylvain Costes, Som Dutta, Talayeh Razzaghi, and Mohammed Eslami, for critical discussion of the analyses.

Abstract

Efforts to mitigate the spread of coronavirus disease 2019 (COVID-19) in the United States require an accurate understanding of how the epidemic is progressing. Datasets available from the National Center for Health Statistics (NCHS) record weekly numbers of deaths attributed to a set of 'select causes' from 1 September 2019 through 12 August, 2020, including deaths from COVID-19 from 1 February through 12 August, 2020 in the entire United States (US), by state, and cumulatively for individual counties. Comparing US and state level deaths from select causes recorded in 2020 with values from 2014-2019 identified a number of changes that exceeded 95% confidence limits on historical mean values, including three states with possible deaths from COVID-19 in December 2019. Comparing the NCHS datasets with data compiled by four public pandemic tracking sites on deaths from COVID-19 suggests that a large number of deaths counted by the public data tracking sites have not yet been reported to the NCHS. Estimates using the percentage of deaths from COVID-19 relative to all Natural Causes as reported to the NCHS and the numbers of COVID-19 deaths counted by the public tracking sites suggests that perhaps 20% of deaths from Natural Causes, as many as 200,000, may not yet have been reported to the NCHS. Evaluating changes in the fractions of deaths attributed to COVID-19 and other specific causes or nonspecific outcomes during the epidemic, relative to 2020 totals or historical mean values, can provide a valuable perspective on the public health consequences of COVID-19.

Main Text

Introduction

The COVID-19 pandemic is disrupting nearly all aspects of global society, with widespread health, economic, and sociocultural impacts from both the disease and responses to it. To ensure that pandemic responses are effective but not extreme, it is essential to characterize how the disease is spreading accurately, and as promptly as possible. Predictive models are valuable tools to inform pandemic response, and a large number of COVID-19 models have been developed (1). Unfortunately, published models have not accurately represented US spread, with highly variable outputs from different models, as well as from updates to a single model (2).

Models can only be as accurate as the data they analyze, and it has been suggested that reported COVID-19 death count data may be more reliable, as inputs to models, than data on positive tests or cases (3). However, this relies on deaths due to COVID-19 being identified accurately as such. Weinberger et al. (4) estimate that ~28% of excess US-wide deaths between March & May were identified as something other than COVID-19, and they report 'substantial variability between states in the difference between official COVID-19 deaths and the estimated burden of excess deaths.' Different datasets produced by the National Center for Health Statistics report inconsistent numbers of deaths from COVID-19, which in some cases lag considerably behind deaths from COVID-19 as tracked by public websites. As well, many comments in the media have raised questions regarding the accuracy of COVID-19 death reporting in the US (e.g., 5).

Recent reports on regression modeling of 'excess deaths' associated with the COVID-19 epidemic in the US suggest that 87,001 deaths in March and April (6), and 122,300 deaths between March and May (4) would not have happened under prior-year rates of death. On 31 July, the New York Times noted 179,800 excess deaths in the US between 15 March and 11 July, of which 45,300 were not attributed to COVID-19 (7). The 19 August release of the Centers for Disease Control (CDC) dataset 'Excess Deaths associated with COVID-19' (8) provides estimates of between 135,593 and 235,728 excess deaths due to the pandemic, depending on how weighting factors are applied.

Regression models generally do not use or provide information about the ratios between different causes of death, nor how these ratios may vary over time. To address this gap, we evaluated the percentages of deaths reported as from different causes each week for consistency over time, and compared these to data on deaths from COVID-19 recorded by four public pandemic tracking sites including the New York Times (7, 9), USA Facts (10), the Johns Hopkins CSSE dashboard (11) and the Atlantic COVID Tracking Project (12).

Here we report seasonal and epidemic-related variability in a subset of weekly reported death ratios, demonstrating that evaluating the ratios of deaths attributed to select causes over time can identify causes of death that are being reported at anomalous rates relative to historical norms. Ratios can also be used to estimate numbers of total deaths based on publicly-tracked numbers of COVID deaths, which could mitigate some aspects of delayed reporting.

Results

CDC Datasets

The National Center for Health Statistics (NCHS) datasets on "Weekly Counts of Deaths by State and Select Causes", in final form for 2014-2018 (13) and as provisional data for 2019-2020 (14), contain weekly counts of deaths as MMWR weeks, starting from the first day of the first year in the dataset, for each of the 50 states, including Puerto Rico and the District of Columbia, with New York City reported separately from New York state. The dataset also reports data for the 'United States' that include all 50 states and the District of Columbia, but not Puerto Rico or other outlying US territories. Entries report the number of deaths, aggregated as 'All Cause' or 'Natural Cause' deaths ('natural' means not caused by some external event such as an accident or deliberate action), as well as a set of 'Select Causes' including Septicemia A40-A41; Malignant neoplasms C00-C97; Diabetes mellitus E10-E14; Alzheimer disease G30; Influenza and pneumonia J09-J18; Chronic lower respiratory diseases J40-J47; Other diseases of the respiratory system J00-J06,J30-J39,J67,J70-J98; Nephritis, nephrotic syndrome and nephrosis N00-N07,N17-N19,N25-N27; Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified R00-R99; Diseases of heart I00-I09,I11,I13,I20-I51; Cerebrovascular diseases I60-I69, COVID-19 U071, Multiple Cause of Death, and COVID-19 U071, Underlying Cause of Death. Hereafter, we refer to codes C00-C79 as 'Cancer'; I00-I09 as 'Heart Disease'; I60-I69 as 'Stroke'; the J codes other than J09-J18 combined as 'Respiratory Illnesses'; and R00-R99 as 'Abnormal Findings'.

The technical notes associated with the CDC and NCHS datasets indicate that data in recent weeks may be less than 25% complete, but are generally at least 75% complete within 8 weeks after the death occurred (15). Notes in the dataset on Excess Deaths Associated with COVID-19, starting with the 1 Feb 2020 entries, state "Data in recent weeks are incomplete. Only 60% of death records are submitted to NCHS within 10 days of the date of death, and completeness varies by jurisdiction. Weights may be too low to account for underreporting, numbers of deaths are likely underestimated." (16).

Weekly US Deaths from 'Select Causes' including COVID-19

The 'Select Causes' in the CDC datasets do not comprise all deaths from 'Natural Causes', so the number of deaths attributed to natural causes other than the listed 'Select Causes' must be obtained by subtraction. Since 2014, using the 'United States' data values for illustration, the number of select cause and unidentified deaths over time show a clear seasonal periodicity, with the majority of the winter increase being associated with deaths from respiratory causes (Fig. 1a), including COVID-19 in 2020. Despite the fluctuating numbers, the ratios of deaths from different causes remain fairly constant over time, until the increase in deaths from COVID causes compression of the other ratios (Fig. 1b). However, when COVID is removed from the assessment, the ratios of deaths from most other causes do not appear dramatically altered, suggesting that most deaths from causes other than COVID-19 are being reported consistently. The only obvious change in the ratios of deaths from different causes is an increase in deaths attributed to Abnormal Findings, in parallel with reductions in the fraction of deaths from Heart Disease, Cancer, and Respiratory Illnesses (Fig. 1c).

Comparing the percentages of individual causes of death per week in 2020 with the mean percentage of weekly deaths during matched MMWR weeks over the period 2015-2019, again for the 'United States' (Fig. 2), facilitates evaluation of the extent to which deaths in 2020 have

been attributed consistently to the individual select causes, relative to historical data. Raw numerical weekly data from 2020 (Fig. 2, thick colored lines) are considered significant when they fall more than two standard deviations away from mean values for the matched MMWR weeks in 2014-2019. Mean values from three-week centered rolling windows of 2020 data (Fig. 2, thin colored lines and shading) are considered significant when separated by more than one standard deviation from the one standard deviation bound on historical data.

Two causes of death show statistically-significant increases relative to prior years, when the fractions of deaths in 2020 are calculated as a percentage of deaths from all Natural Causes in 2020 including COVID-19 (Fig. 2, thick dashed lines): Abnormal Findings has increased monotonically since the start of the pandemic (Fig. 2a), which confirms the apparent increase visible in Fig. 1; and Influenza and Pneumonia increased significantly above the prior-years mean only during the month of March (Fig. 2b). The fractions of deaths from all other causes drop significantly below historical mean values, when calculated as a percentage of all deaths from Natural Causes including COVID-19 (Fig. 2 c-k). Again, this apparent reduction in the percentages of deaths from causes other than COVID-19 is an artefact of including deaths from COVID-19 in the number of Natural Cause deaths used to calculate these ratios, because deaths from COVID-19 represented at least 10% of all weekly deaths from March through May (Fig. 1b) and the historical data had no deaths from COVID-19.

Several different patterns emerge when the fractions of deaths in 2020 are calculated after subtracting deaths from COVID-19 from 2020 Natural Cause deaths (Fig. 2, dash-dotted lines). The fractions of deaths from Cancer (Fig. 2c) and Respiratory Illnesses (Fig. 2d) remain significantly below the mean from prior years, except for a sharp inflection in the most recent Respiratory Illnesses datapoint, that is likely caused by incomplete recent data. The fractions of deaths from Nephritis (Fig. 2e) and Stroke (Fig. 2f) cease to show any significant difference from prior-year mean values. The fractions of deaths from Heart Disease (Fig. 2g) and Septicemia (Fig. 2h) show a non-significant uptick in April, and then drop significantly below the prior-year mean values. The fractions of deaths from Diabetes (Fig. 2i), Alzheimer's (Fig. 2j) and the residual Unidentified deaths not attributed to a specific 'select cause' (Fig. 2k) show significant increases above prior year mean values in March and April, and then return towards mean values, showing no significant differences from historical mean values by June.

However, these ratios still may be inappropriate to compare with historical data, because the fraction of deaths attributed to Abnormal Findings increased so dramatically above historical values in 2020 (Fig.s 1c and 2a), with each of Abnormal Findings and COVID-19 accounting for approximately 6% of deaths in June (Fig. 1b). To perform a true comparison of historical values with data from 2020, deaths from Abnormal Findings should also be excluded from the calculations (Fig. 2, thick solid lines). Only the fraction of deaths from Respiratory Illnesses still shows a significant reduction (Fig. 2d), when the percentages of deaths in 2020 are calculated subtracting both COVID-19 and Abnormal Findings from the denominator. Deaths from Influenza (Fig. 2b), Cancer (Fig. 2c) Nephritis. (Fig 2e), Heart Disease (Fig. 2g) and Sepsis (Fig. 2h) now show no significant differences from historical values. Deaths from Stroke (Fig. 2d), Diabetes (Fig. 2i), Alzheimer's (Fig. 2j) and deaths not otherwise identified (Fig. 2k) now join deaths from Abnormal Findings in showing significant increases in the percentages of deaths attributed to these causes in 2020 relative to historical values. Although deaths from Abnormal Findings were not subtracted from the historical mean data, this category of deaths historically

represented ~1.2% of total deaths (Fig. 2a), which is within the prior-year standard deviation for causes of death that occur at larger percentages.

The most recent few weeks of data for a number of the 'Select Causes' show inflections away from the curve traced by older timepoints. In weekly CDC dataset releases going back to May, the most recent two weeks of data usually provide unreliable ratios of deaths (data not shown). Data older than 2 weeks from the date of release provide percentages of deaths that are accurate based on values calculated from later datasets, despite residual incompleteness of the reported data. These observation confirms that even incomplete data can provide accurate information about the ratios between different causes of death.

Weekly Deaths Reported by State

At the level of the entire US, the ratios of reported deaths from nearly all the 'select causes' are significantly different from historical values at 95% confidence or greater, for at least some weeks during 2020. It is also useful to examine death reporting at state level, because the CDC notes that 'completeness varies by jurisdiction' (16). The ratio of the number of weekly deaths in 2020 divided by 2014-2019 average number of deaths from the corresponding MMWR week is a normalized indicator that allows comparison between states with widely-varying populations. To facilitate examination of significant differences, in Figure 3 points that are within 95% confidence limits of historical mean values, or that were calculated with imputed rather than reported values, have been removed.

Across the majority of states, deaths from Abnormal Findings have been increasing almost monotonically since Feb. 2020 (Fig 3a, recent weeks indicated by larger points). Curiously, since mid-May, Louisiana and Minnesota have simultaneously reported increases in the percentage of deaths from Abnormal Findings and decreases in the percentage of deaths from COVID-19 with Pearson correlation coefficients lower than -0.35, although most states and the entire US show positive Pearson correlations between percentages of deaths from Abnormal Findings and COVID-19 (Fig. 3a, shading from blue to green as points get larger and higher up the plot).

In contrast, since February only a few states have reported more than twice the historical levels of deaths from Influenza and Pneumonia (Fig. 3b), most of which were associated with high levels of COVID-19 deaths. Only Arizona has reported recent increases in deaths from Influenza and Pneumonia, and these are also associated with increases in reported deaths from COVID-19 (Fig. 3b, crosses shading from blue to green as points get larger and higher up the Y axis). Interestingly, a number of states are reporting significantly lower levels of pneumonia than historical norms, some of which are also associated with higher levels of COVID-19 (Fig. 3b, blue vs. green points with values below 1). Deaths from other respiratory illnesses show similar patterns as Influenza (Fig. 3d).

Deaths from Cancer (Fig. 3c), Stroke (Fig. 3f), and Septicemia (Fig. 3h) show relatively small changes from historical values, and no consistent correlation with deaths from COVID-19. The majority of states reported significantly fewer deaths from cancer than in prior years (Fig. 3c, values below 1).

For the most part, deaths from Heart Disease (Fig. 3g), Diabetes (Fig. 3i), and Alzheimer's (Fig. 3j) are all reported at higher levels in association with COVID, although a larger number of states

are reporting significantly lower levels of Heart Disease than in prior years (Fig. 3g, points below one), than for other causes of death. Reported deaths from Nephritis (Fig. 3e) mostly show similar correlations with COVID-19 deaths, although Arizona recently reported elevated levels of nephritis with low levels of COVID-19. Deaths from causes not identified in the CDC dataset (Fig. 3k) also show correlations with deaths from COVID-19 in most states, except for New Hampshire, Rhode Island, and Idaho, which are reporting anomalously high numbers of deaths from unidentified causes without reporting deaths from COVID-19.

The majority of states reported significantly low levels of deaths from cancer, and multiple states reported significantly low levels of deaths from other select causes of death. A number of states reported fewer than half the mean historical number of deaths from all Natural Causes excluding COVID-19 (Fig. 3l). The majority of these points are from weeks in July and August, and are therefore likely to be a result of incomplete reporting, but the numbers of deaths recorded for at least one week in February and March are still below 90% of historical numbers of deaths from Natural Causes in the corresponding MMWR week for the localities Idaho, Oklahoma, Pennsylvania, Puerto Rico, Rhode Island, Utah, Vermont, and Wyoming.

Before pursuing the question of anomalously low numbers of reported deaths, it is worthwhile to address the anomalously high numbers of reported deaths from Abnormal Findings, as seen in Fig. 2a extending back prior to 2020. At state level, the three states California, New York, and North Carolina reported deaths from Abnormal Findings that exceeded 95% confidence limits for historical mean values for the last two weeks of 2019, and also reported more than three times the historical mean number of deaths from Abnormal Findings for at least two weeks in January 2020 (Fig. 3m).

Cumulative Deaths from COVID-19 by County

In 2020, individual states reported statistically-significant differences in the numbers of deaths from select causes (Fig. 3) that certainly contribute to the percentages reported at the level of the entire US (Fig.s 1 and 2), but deaths are typically reported at the level of a county coroner. County-level historical data on monthly deaths by county are available from the CDC Wonder database (17), and data on deaths from COVID-19 and deaths from All Causes are available in the CDC dataset reporting 'Provisional COVID-19 Death Counts in the United States by County' (18). In the 26 August release of this dataset, 923 of the 3147 US counties are recorded as having at least one death from COVID-19. Comparing these data to county-level counts of death from COVID-19 as collected by the New York Times (9); USA Facts (10); and the Johns Hopkins CSSE dashboard (11) identifies significant anomalies at county level between current and historical counts of deaths from All Causes and deaths from COVID-19 (Fig. 4).

County-level ratios of CDC-recorded deaths from All Causes in 2020 over 2014-2018 mean values (Fig. 4a) display large anomalies in several states. Six counties in Virginia, four counties in Mississippi, two counties in New York, and one county each in Arizona, Georgia Iowa, Kentucky Missouri, Pennsylvania, South Carolina and West Virginia reported more than twice as many deaths from All Causes over the period 1 February - 26 August 2020, cumulatively, than historical mean values (Fig. 4a) -- an increase of over 100%. However, only the two counties in New York (the Bronx and New York proper) and Manassas City in Virginia reported more than 20% of these deaths as being from COVID-19 (Fig. 4a, yellow/orange colors). Of the 555 counties reporting numbers of total deaths in 2020 that exceed 95% confidence limits on prior-

year mean values, 373 of them reported fewer than 10% of these deaths as being caused by COVID-19 (Fig. 4a, blue and dark green points higher up the Y axis).

Some counties reporting total numbers of deaths within historical mean 95% confidence limits still report large percentages of deaths from COVID-19 (Fig. 4a, large points in warm colors near 1 on the Y axis). For example, Jenkins County in Georgia reported the two more deaths in 2020 as the mean from prior years, yet also reported 25% of 2020 deaths in that county were from COVID. Of the 325 counties that reported fewer total deaths in 2020 than 95% confidence limits on prior year mean values, 87 of them also reported at least 10% of those deaths as being from COVID.

Further anomalies are observed when the CDC county-level death data are compared with county-level death data recorded by the public COVID-19 tracking sites. Although only 877 counties have reported COVID-19 deaths to the CDC, as of 19 August at least one COVID-19 death had been recorded in 3197 counties by the New York Times; 3144 counties by the organization USA Facts; and 3221 counties by the Johns Hopkins CSSE. Cumulative death counts calculated from the daily data maintained by these three public tracking organizations are very similar (e.g., Fig. 4b), with the majority of the discrepancies related to counties around New York City. However, weekly deaths as recorded by the CDC show considerably more scatter when compared with each of the public tracking project datasets, with the CDC recording more deaths in some counties and fewer deaths in others (e.g., Fig. 4c).

The public data trackers have recorded deaths from COVID-19 in every county that reported at least one death to the CDC. The NYT, for example, has recorded more deaths than the CDC in 535 counties, for which the ratios range upwards from 1 to 7 (Fig. 4d, colored points above 1). In most cases, the number of COVID-19 deaths reported to the CDC is a small percentage both of total deaths in 2020 (point color) and of prior year mean values (point size) recorded by the CDC. Additionally, the NYT recorded at least one COVID death in 1530 counties where the CDC records no COVID-19 deaths (Fig. 4d, small grey points): in 259 of these counties the NYT recorded between 10 and 56 deaths from COVID-19 since Feb. 2020, although the 26 August release of the CDC dataset on "Provisional COVID-19 Deaths by County' has recorded none.

In the 26 August dataset release, the CDC had recorded more deaths than the USA Facts dataset in 5550 counties, and 83 of these counties have reported at least twice as many deaths from COVID to the CDC as are recorded by USA Facts (Fig. 4e, colored points above 1). There are 689 counties where no deaths have been recorded either by the CDC or the public data trackers (Fig. 4e, small grey points at 10 on the Y axis): this includes counties from nearly all states.

Changes in COVID-19 Death Count Data over Time

These data on cumulative deaths from COVID-19 at county level paint a concerning picture regarding the extent to which deaths from COVID-19 are being reported accurately, but it is possible that local anomalies aggregated at the level of individual states or the entire US balance each other out. For the most part, public COVID-19 tracking site datasets released on different dates do not show major discrepancies in the numbers of deaths recorded, except for differences possibly associated with whether data revisions were back-distributed to prior dates or made on the date of announcement (data not shown). Curiously, considering the discrepancies in cumulative county-level data just presented, the dataset on 'United States

COVID-19 Cases and Deaths by State over Time' (19) reports daily death data at state level that is nearly identical to one or other of the public trackers, although which one varies by entry.

In contrast, and consistent with CDC notifications, recent releases of the CDC's 'Select Causes' dataset record larger numbers of deaths going back to March and April, even compared to the release dated 15 July. For example, for the week ending 18 April, the 15 July release recorded 15,446 deaths from COVID-19 in the 'United States' (not including Puerto Rico or the other outlying territories), while the 22 July release recorded 16,060 deaths, and the 26 August release records 16,079 (data not shown). The 26 August release of the 'Select Causes' dataset retains the same data as the 19 August dataset for every week prior to the week ending 12 August, indicating that some updates to this dataset are made to the most recent week only.

Comparing data reported in the 21 May and 26 August releases of the CDC 'Select Causes' dataset with death counts from the four public COVID-19 tracking sites provides a view of changes over time (Fig. 5). Early in the pandemic, there was significant uncertainty about how to classify deaths from COVID-19, and public tracking activities were new -- even so, the 21 May release of the CDC 'Select Causes' dataset recorded relatively similar numbers of deaths from COVID-19 to those from the public tracking sites, across states as well as in the US as a whole (Fig. 5, '21 May'). However, additional deaths from COVID-19 have continued to be reported to the CDC, and the 26 Aug dataset records more deaths in the same set of MMWR weeks 11-16 than the public tracking sites do, for nearly every state as well as the US (Fig. 5, '19 Aug.', '11-16'). In May and June, the public tracking sites counted more deaths from COVID-19 than are recorded in the 26 August CDC dataset, in the US as a whole as well as most states (Fig. 5, '19 Aug.', '17-21' and '22-26'). In July, for the US as a whole, the public tracking sites recorded similar numbers of deaths as reported in the 26 Aug. CDC dataset release, but this masks considerable variation at the level of individual states, as to whether the CDC or the public tracking sites recorded more deaths from COVID-19 (Fig. 5, '26 Aug.', '27-31').

Estimating Missing Deaths

The CDC records and releases death information in the 'Select Causes' datasets only as it is reported to them, while the public tracking sites are often considerably more prompt in publicizing COVID-19 death counts. Because both sets of data describe the same results, i.e., the number of people who have died from COVID-19, it should be possible to combine information from the CDC about the ratios of deaths from different causes with the numbers of deaths as recorded by the public data tracking sites, to estimate the total number of deaths from Natural Causes that will eventually be reported to the CDC. Estimates are based on the number of deaths from COVID-19 reported by the public tracking sites, using the identity:

(1)

$$\frac{\text{COVID}_{\text{CDC}}}{\text{NaturalCause}_{\text{CDC}}} = \frac{\text{COVID}_{\text{Tracker}}}{\text{NaturalCause}_{\text{Tracker}}}$$

If the hypothesis is correct that data from both the CDC and public tracking sites are reliable, then dividing the number of deaths from COVID-19 reported by each of the public tracking sites with the percentage of deaths from COVID-19 over all Natural Causes reported in the CDC

'Select Causes' dataset, should give the expected number of deaths from natural causes for the tracked number of COVID-19 deaths, which can be subtracted from the recorded value to obtain the difference between the estimate and reality (Table 1). The negative differences for MMWR weeks 12-16 over the entire US suggest that the number of COVID-19 deaths recorded early in the pandemic, by both the public tracking sites and the CDC, represented a considerable under-count of actual deaths from COVID-19.

The 26 August dataset gives considerably larger deficits in the estimates for MMWR weeks 12-16 than the 21 May dataset, which is due to the greater completeness of the later CDC dataset for deaths in March and April. Excluding already-negative numbers, for MMWR weeks 17-31 there could be from 129,876 to 202,570 deaths from Natural Causes that have not yet been reported to the CDC, with an average of 172,712.

Discussion

The US Dept. of Health and Human Services' *COVID National Diagnostic Strategy* (20) recommends using deaths from COVID-19 as reported to the CDC as one criterion to inform policy decisions on reopening. A number of online tools evaluating the current status of reopening readiness also use CDC-reported death data as inputs (e.g., 21). Evaluating the completeness of reported deaths from COVID-19 is an important aspect of strategies for ensuring that the US reopens safely. Further, assessing the extent of mis-attribution of deaths between COVID and other causes, as well as possible under-reporting, is a key aspect of understanding the accuracy of death data for use in epidemiological models.

Fractions of Death

During an epidemic, deaths from causes not affected by the epidemic disease should continue occurring at similar rates during the epidemic as before, meaning the ratios between these causes of deaths would be preserved, even when the reported numbers of deaths are not complete, after some threshold level for accuracy of reporting has been met. A metric for evaluating the consistency of death reporting during an epidemic, therefore, is the extent to which percentages of deaths from specific causes not associated with the epidemic, taken as a fraction of total deaths from causes other than the epidemic, remain within historical norms. If deaths are being reported consistently, then the percentages of individual non-epidemic causes of death should remain within historical 95% confidence limits, as a fraction of deaths from non-epidemic natural causes.

An overview of deaths in the entire US since January 2014 demonstrates that ratios of different deaths remained relatively constant, before the COVID-19 pandemic, aside from minor shifts in the ratios of respiratory diseases and cancer that are associated with winter prevalence of influenza (Fig. 1). The introduction of COVID-19 produced an apparent reduction in the percentages of other causes of death (Fig. 1b), which could explain claims circulating on social media that deaths from other causes were being coded as from COVID-19. However, this apparent reduction is an artefact caused by including the additional deaths from COVID-19 in the total deaths from Natural Causes when calculating the other percentages: subtracting COVID-19 returns almost all the others nearly to pre-pandemic levels (Fig. 1c).

In 2020, the ratios of nearly all natural causes of death, aside from deaths recorded as COVID-19 and in the R00-99 codes for Abnormal Clinical Findings, show excursions of a few percent of total natural cause deaths in March and April (Fig. 2), likely associated with mis-attribution of deaths associated with COVID-19 to these causes. Nearly all natural causes of death returned to within one percent of historical mean values by May, although a highly significant increase in the percentage of deaths reported under the R00-R99 codes (Fig. 2a), and a smaller but still significant increase in deaths not identified as one of the 'select causes' (Fig. 2k), may indicate ongoing attribution of deaths associated with COVID-19 to these causes.

Several small yet significant shifts in the US-wide percentages of other deaths also have interesting implications. The significant reduction in the percentage of deaths attributed to respiratory illnesses other than Influenza and Pneumonia (Fig. 2d) may indicate that people with these illnesses are more susceptible to SARS CoV-2 infection and their deaths are being attributed to that cause, although their deaths might have been attributed to other respiratory causes had to the COVID-19 pandemic not occurred. Conversely, the significant increases in the percentages of deaths attributed to Diabetes (Fig. 2i) and Alzheimer's (Fig. 2j) may indicate that SARS CoV-2 infection is elevating mortality for people with these diseases, yet their deaths are being attributed to their underlying disease.

Although the percentage of deaths from Heart Disease increased significantly in April during an early peak of the pandemic (Fig. 2g), the percentage subsequently returned to historical levels, which could mean that cardiac patients who die with COVID-19 are being coded accurately as dying from COVID-19, or alternatively that SARS CoV-2 infection does not exacerbate cardiac diseases as much as some assessments have reported (22,23). In contrast, the percentage of deaths attributed to Stroke was not significantly elevated during the April peak (Fig. 2f), but has since increased steadily and is now significantly higher than historical levels, which is consistent with reports that SARS CoV-2 infection causes and/or increases the severity of coagulation disorders (24).

The percentages of the various causes of death at the level of individual states paint a much more complicated picture (Fig. 3). Although many states contribute to the US-wide increase in reported deaths from Abnormal Findings, several states show very little increase (Fig. 3a). The US-wide recent reductions in Respiratory Illnesses (Fig. 3b), including Influenza and Pneumonia (Fig. 3d), are replicated in many states, although increases in both, associated with the April COVID-19 peak in New York and New Jersey, are also clear. Similarly-associated increases in Nephritis (Fig. 3e), Heart Disease (Fig. 3g), Diabetes (Fig. 3i), and Alzheimer's (Fig. 3i) are apparent, as are increases from Septicemia (Fig. 3h) and deaths from Unidentified causes (Fig. 3k). At minimum, this indicates considerable inconsistency between different states in how deaths from various causes are being reported.

More recent increases in Nephritis (Fig. 3e), Sepsis (Fig. 3h), Diabetes (Fig. 3i), Alzheimers (Fig. 3i), and Unidentified causes (Fig. 3k) are confined to only a few states. Considering reported recent increases in SARS CoV-2 infection in Arizona (7), the recent increases in Influenza and Pneumonia (Fig. 3b), Cancer (Fig. 3c) and Stroke (Fig. 3f) reported by that state could indicate mis-attribution of deaths from COVID-19.

Despite a few states reporting significantly increased numbers of deaths from individual causes, the most obvious result from comparing current to historical mean numbers of death is that

multiple states report numbers of deaths from all Natural Causes not including COVID-19 that are significantly lower than prior year values (Fig. 3l). This suggests that extended delays in reporting could have a considerable impact on the accuracy of deaths counts recorded by the CDC, a concern that is further strengthened by the observation that a considerable number of counties have reported significantly fewer deaths between February and July 2020, than historical mean values for the same period over prior years 2014-2018 (Fig. 4a).

Estimating Deaths in the Context of COVID-19

At county level, considering only the CDC 'Select Causes' dataset, it is quite clear that deaths reported as due to COVID-19 do not correlate with reported numbers of deaths from all causes (Fig. 4a). Comparing data recorded by the CDC to deaths from COVID-19 as counted by the public data trackers (Fig. 4d and e) emphasizes even more strongly that multiple hundreds, and perhaps thousands, of deaths probably from COVID-19 are not yet recorded as such by the CDC. The CDC acknowledges that delays in reporting of deaths are much increased during the pandemic (8, 14-16, 18, 19), and it is to be hoped that many of these deaths are currently in the pipeline for being reported to the CDC. Unfortunately it is also possible that some paperwork is remaining un-processed, particularly in counties with limited public health infrastructure, as a consequence of the current extraordinary strain on US public health systems.

As we have shown, after the third week of reporting the ratios of deaths from different causes recorded by the CDC accurately represent the eventual outcome (Fig. 2). This recognition enables a novel approach that can compensate for delays in reporting deaths to the CDC, by taking advantage of the rapid availability of COVID-19 death counts as collated by the public tracking sites. Estimates of US-wide deaths from COVID-19 based on counts from public tracking sites are too low, early in the pandemic (Table 1), likely due to inadequate testing and discrepancies in how deaths were reported (e.g., the debate over 'provisional' vs. 'confirmed').

After CDC guidance on counting deaths from COVID-19 was released on 14 April (25), deaths should have been reported more consistently, and the estimates for still-unreported deaths from late April through May (MMWR weeks 17-21) are relatively modest, at less than 15% of the current CDC total (Table 1). For June (MMWR weeks 22-26), in contrast, the estimated number of deaths that are not-yet-reported is almost 40% of the current CDC total, despite the CDC already recording more deaths than the public data tracking sites in several states, including the large states Texas and Florida (Fig. 5). In July (MMWR weeks 27-31), although the CDC data are still highly incomplete for these recent weeks, the estimate of missing deaths from COVID-19 death counts compiled by the public tracking sites are barely 10% of deaths already recorded by the CDC (Table 1). The CDC reports that death data are only 75% complete after 8 weeks under normal circumstances (15), so this low estimate suggests anomalies in these data. One possible explanation is that a large number of deaths from COVID-19 were somehow not counted by the public tracking sites. Considering the high percentage of deaths from COVID-19 reported in July and August (Fig.s 3 and 4), a more probable explanation is that the CDC has started recording deaths from COVID-19 much more rapidly than deaths from other causes and therefore the ratio identity is no longer correct.

Our estimates for total deaths from natural causes using the weekly percentages of deaths from COVID-19 reported to the CDC, combined with the weekly counts of deaths from COVID-19 compiled by public data trackers, suggest that 21% of deaths from natural causes over the

period May-July 2020 have not yet been reported (Table 1). At state level and for weeks early enough for the CDC data to be substantially complete, COVID-19 death counts from the public tracking sites are lower than CDC-reported numbers (Figure 5), which suggests it is likely that even these appalling numbers are under-estimates.

Limitations and Open Questions

Analyses are only as good as the data used to make them, and the data available on deaths from COVID-19 in the US are highly unreliable. This report has focused on exploring discrepancies in available data on deaths from COVID-19 and other causes during these extraordinary times, and it seems unlikely, unfortunately, that available data exaggerate the impact of COVID-19. Public data trackers have counted deaths from COVID-19 in twice the number of counties as have reported COVID-19 deaths to the CDC. Many counties, even while reporting deaths from COVID-19 to the CDC, at the same time reported numbers of deaths from all natural causes in 2020 that are significantly lower than the mean historical values for the corresponding MMWR weeks in 2014-2018 (Fig. 4a). The US population is higher in 2020 than in prior years, so the observation that recorded deaths in so many counties are below the lower 95% confidence limits for historical data is even more improbable than statistics might suggest.

A further anomaly in death data recorded by the CDC is the unprecedented increase in deaths coded as R00-R99, 'Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified'. Sometimes, deaths are recorded under the R codes only temporarily, until the actual cause of death can be determined and the record updated (26). This does not entirely explain the observed increase, however, because a CDC data release listing only deaths in code R99, "Unknown Underlying Cause of Death" (27), contains state-level numbers that mirror the increases in deaths reported under the aggregated R00-R99 codes.

One plausible explanation for this anomalous increase is that deaths actually caused by COVID-19 are being reported as Abnormal Findings, when classification criteria for other causes of death were not met. If true, then the actual numbers of deaths from COVID-19, since June, could be up to half-again higher than has been recognized, because at least 5% of total US deaths from natural causes since that time were reported under the R00-99 codes (Fig. 1c).

Early increases in deaths from Abnormal Findings may hint at answers to another vexing question about SARS-CoV-2, namely when the virus started to spread in the US. Three states (California, New York with New York City, and North Carolina) reported rapid and significant increases in deaths from Abnormal Findings over multiple weeks from the end of December 2019 into January 2020 (Fig. 2m). Observations that COVID-19 was circulating in Europe in December 2019 (28, 29) suggest that COVID-19 infection could also be a plausible explanation for these increases. As with the European studies, this hypothesis should be tested by assaying samples from those locations, archived during the relevant timeperiods, for the presence of SARS-CoV-2.

If the winter increase in deaths attributed to the R00-R99 codes for Abnormal Findings is associated with spread of SARS-CoV-2, then unexplained increases in the number of deaths reported under R00-99 codes, or sudden deviations from historical ratios for any causes of death more generally, could be a useful alert trigger for public health surveillance.

Materials and Methods

Data preprocessing

Data on "Weekly Counts of Deaths by State and Select Causes, 2019-2020" (https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-State-and-Select-Causes/muzy-jte6) were downloaded weekly since 21 May, 2020. The 26 August, 2020 releases of the above datasets were used to generate the figures and tables presented here. In addition, the 5 June 2020 release of data on "Weekly Counts of Deaths by State and Select Causes, 2014-2018" (https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-State-and-Select-Causes/3yf8-kanr) were used for all analyses presented here. State and county-level data on deaths from COVID-19 were obtained from four public pandemic tracking sites: the New York Times (9); USA Facts (10); and the Johns Hopkins CSSE dashboard (11); and state-level data from the Atlantic COVID Tracking Project (12), using the 26 August release.

The 8 July release of the NCHS datafile on Select Causes temporarily omitted numbers for the US as a whole. In the process of trying to regenerate these numbers, we compared the all-US data entries for 'Total Deaths' in several 'Provisional COVID-19' datasets with the 'All Cause' and 'Natural Causes' deaths in several 'Select Causes' datasets archived from May and June. Even in data more than two months prior to the download date, a number of inconsistencies of several hundred to a few thousand total deaths were observed between the two data sources, that could not be explained by the inclusion/exclusion of numbers for Puerto Rico. It is possible that additional death information from other US territories, to which we did not have access, might explain the differences. This observation led us to compare individual cause of death data only to other data presented in the same CDC dataset.

A major objective of the current work is to include all states and available data in the analyses presented. However, the NCHS suppresses data entries with numbers of deaths between 1 and 9. Because of the above-mentioned inconsistencies in total death data, the application of algorithms to allocate the small numbers of suppressed death counts by subtracting known numbers of deaths from reported totals had the potential to propagate low-confidence information. Instead, for all datasets, suppressed values for deaths from COVID-19 were replaced with '1's, to take the lower bound on counts of deaths associated with COVID-19. In the "Weekly Select Causes" datasets, suppressed values for causes of death other than COVID-19 were replaced by '9's, in order to minimize the effect of increases in the numbers of 'select causes' of death by using the upper bound estimate.

Initial analyses were performed on three datasets, in which all 1s or all 9s were inserted in addition to the above, but only the 'replace by 9/1' results are reported here. This approach for estimating suppressed counts of deaths is chosen because observations that are statistically significant under these assumptions could only become more significant if more correct numbers, which would be smaller for 'Select Cause' deaths and larger for COVID-19 deaths, were used.

Analysis

The programs Orange Data Mining (30) and Microsoft Excel (Microsoft Corp.) were used for data preprocessing and analysis. After downloading each dataset, empty cells were imputed with either a 1 or a 9 to generate 'lower bound' and 'upper bound' datasets. The 'Unidentified' deaths column was generated by subtracting all the individual 'select cause' columns including only 'COVID-19 as an Underlying Cause', from the 'Natural Cause' column. The Respiratory Illness column was generated by combining the 'Chronic lower respiratory diseases (J40-J47)' and the 'Other diseases of respiratory system (J00-J06, J30-J39, J67, J70-J98)'. These two causes of death were combined because they are relatively small, describe closely related illnesses, and fluctuate similarly.

Percentage values were calculated by dividing the individual cause of death columns in the 'Select Causes' datasets by the 'Natural Causes' columns, or the 'Natural Causes' with deaths from 'COVID-19 as an Underlying Cause' or both Underlying COVID-19 and Abnormal Findings deaths subtracted. In some recent-week data, subtracting deaths from COVID-19 and/or Abnormal Findings from Natural Cause deaths produced 0s, so we added 1 to the in the denominator when subtractions were performed. 3-week and 5-week trailing averages and standard deviations were calculated using the 'Moving Transform' widget in Orange Data Mining, and 3-week centered moving averages were also calculated using Excel, as a check on the trailing-average calculations. The use of trailing averages shifts the features of the curves by half the trailing-average window, so Python was used to re-calculate and plot both matched MMWR week and centered moving-window averages, with 1 and 2 standard deviation limits indicated by shading.

Differences in death counts recorded in the 21 May and 26 August releases of the CDC 'Select Causes' dataset as compared to the public tracking sites were determined by summing the data in five-week bins spanning MMWR weeks 12-16, 17-21, 22-26 and 27-31, covering the time period between 14 March 2020 and 8 August 2020. Data from the four public tracking sites were binned individually, then averaged to provide consensus values for comparison with the CDC data. The coefficient of variation for each locality was calculated across matched MMWR weeks, as a metric for variability between public tracking sites, and states were considered highly variable if the state-level coefficient of variation exceeded twice the US coefficient of variation of 0.031, for at least 8 weeks in the period. These calculations were performed independently using both Python and Orange Data Mining, and heatmap plots were generated using the pheatmap package (https://cran.r-project.org/web/packages/pheatmap/pheatmap.pdf) in the R programming language (https://www.r-project.org/about.html). Analyses and supplements are available from the Coronavirus International Research Team Github (https://github.com/COV-IRT/dmwg-covid-reporting-qc/).

Statistics

One question addressed in this study is the extent to which the percentages of different classes of deaths are preserved in data from recent weeks where the numbers reported are known to be incomplete. Numbers of deaths follow the Poisson distribution, but the large number of events and conversion to percentages mean the data should be close to normally-distributed. We used running-window averages rather than fitting Poisson regression curves, to make use of the information carried in the variability present in the raw data and identify significantly low as

well as significantly high excursions. Moving averages and standard deviations were calculated for 3 and 5 week running-windows in a single year, with 3-week averages presented because the seasonal trends meant that 5-week moving averages had higher standard deviations than 3-week averages. Single-week averages and standard deviations were calculated individually for each MMWR week across years 2014-2019. Divergences were considered significant when values two standard deviations from the mean for data from centered moving window averages in the winter/spring of 2019-20 ceased to overlap with the 95% confidence interval on the average of matched weekly data from prior winters 2014-15 through 2018-19. Pearson correlation coefficients were calculated for each set of locality data individually using the 'Correlations' widget in Orange.

References

[1] J. T. Unwin *et al.*, "Report 23: State-level tracking of COVID-19 in the United States," no. May, 2020. doi: https://doi.org/10.25561/79231

[2] I. Holmdahl and C. Buckee, "Wrong but Useful — What Covid-19 Epidemiologic Models Can and Cannot Tell Us."

[3] F. S. Lu, A. T. Nguyen, N. B. Link, M. Lipsitch, and M. Santillana, "Estimating the Early Outbreak Cumulative Incidence of COVID-19 in the United States: Three Complementary Approaches.," *medRxiv Prepr. Serv. Heal. Sci.*, 2020.

[4] D. M. Weinberger *et al.*, "Estimation of Excess Deaths Associated With the COVID-19 Pandemic in the United States, March to May 2020," *JAMA Intern. Med.*, vol. 06520, no. May, 2020.

[5] S. Pappas, "How COVID-19 deaths are counted," *Sci. Am.*

[6] S. H. Woolf, D. A. Chapman, R. T. Sabo, D. M. Weinberger, and L. Hill, "Excess Deaths From COVID-19 and Other Causes, March-April 2020," *Jama*, vol. 217, pp. 15–17, 2020.

[7] New York Times"https://www.nytimes.com/interactive/2020/04/21/world/coronavirus-missing-deaths.html".

[8] Centers for Disease Control "https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36".

[9] New York Times "https://github.com/nytimes/covid-19-data".

[10] USA Facts "https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/".

[11] "COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU)." [Online]. Available at: "https://github.com/CSSEGISandData/COVID-19".

[12] "The COVID Tracking Project." [Online]. Available: "https://covidtracking.com/data/api".

[13] National Center for Health Statistics "Weekly Counts of Deaths by State and Select Causes, 2014-2018" [Online]. Available: https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-State-and-Select-Causes/3yf8-kanr.

[14] National Center for Health Statistics "Weekly Counts of Deaths by State and Select Causes, 2019-2020." [Online]. Available: https://data.cdc.gov/NCHS/Weekly-Counts-of-Deaths-by-State-and-Select-Causes/muzy-jte6.

[15] National Center for Health Statistics "Technical Notes on Provisional Death Counts for Coronavirus Disease (COVID-19)." [Online]. Available: https://www.cdc.gov/nchs/nvss/vsrr/covid19/tech_notes.htm.

[16] National Center for Health Statistics "Excess deaths associated with COVID-19." [Online]. Available: https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm.

[17] Centers for Disease Control "Wonder database." [Online]. Available: https://wonder.cdc.gov/.

[18] National Center for Health Statistics "Provisional COVID 19 Death Counts in the United States by County." [Online]. Available: https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-in-the-United-St/kn79-hsxy

[19] Centers for Disease Control "'United States COVID-19 Cases and Deaths by State over Time' ." [Online]. Available: https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36

[20] Dept. of Health and Human Services, "COVID National Diagnostic Strategy," 2020. [Online]. Available: https://www.democrats.senate.gov/imo/media/doc/COVID%20National%20Diagnostics%20Strategy%2005%2024%202020%20v%20FINAL.pdf

[21] "Covid Exit Strategy." [Online]. Available: www.covidexitstrategy.org.

[22] M. M. Riccardo M. Inciardi, Laura Lupi, Gregorio Zaccone, Leonardo Italia, Michela Raffo, Daniela Tomasoni, Dario S. Cani, Manuel Cerini, Davide Farina, Emanuele Gavazzi, Roberto Maroldi, Marianna Adamo, Enrico Ammirati, Gianfranco Sinagra, Carlo M. Lombardi, "Cardiac Involvement in a Patient With Coronavirus Disease 2019 (COVID-19)," *JAMA Cardiol.*, vol. 5, no. 7, 2020.

[23] K. Srivastava, "Association between COVID-19 and cardiovascular disease," *IJC Hear. Vasc.*, p. 100583, 2020.

[24] T. Iba, J. H. Levy, M. Levi, J. M. Connors, and J. Thachil, "Coagulopathy of Coronavirus Disease 2019," *Crit. Care Med.*, vol. Publish Ahead of Print, pp. 1–7, 2020.

[25] Centers for Disease Control "About CDC COVID-19 Data" [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/about-us-cases-deaths.html

[26] "2020 ICD-10-CM Diagnosis Code R99 "[Online]. Available: https://www.icd10data.com/ICD10CM/Codes/R00-R99/R99-R99/R99-/R99

[27] Centers for Disease Control "Mortality-Medical Unknown Underlying Cause of Death (R99), 2017-2020" [Online]. Available: https://www2a.cdc.gov/nchs/vscp/MED_REPORTS/R99_forecast_report.pdf

[28] A. Deslandes *et al.*, "SARS-CoV-2 was already spreading in France in late December 2019," *Int. J. Antimicrob. Agents*, vol. 55, no. 6, p. 106006, 2020.

[29] La Rosa G., M. Iaconelli, P. Mancini, G. Bonanno Ferraro, C. Veneri, L. Bonadonna, L. Lucentini, E. Suffredini. First detection of SARS-CoV-2 in untreated wastewaters in Italy, Science of The Total Environment, Volume 736, 20 Sept. 2020, 139652 https://doi.org/10.1016/j.scitotenv.2020.139652

[30] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: Data Mining Toolbox in Python, Journal of Machine Learning Research 14(Aug): 2349–2353.

Figures and Tables

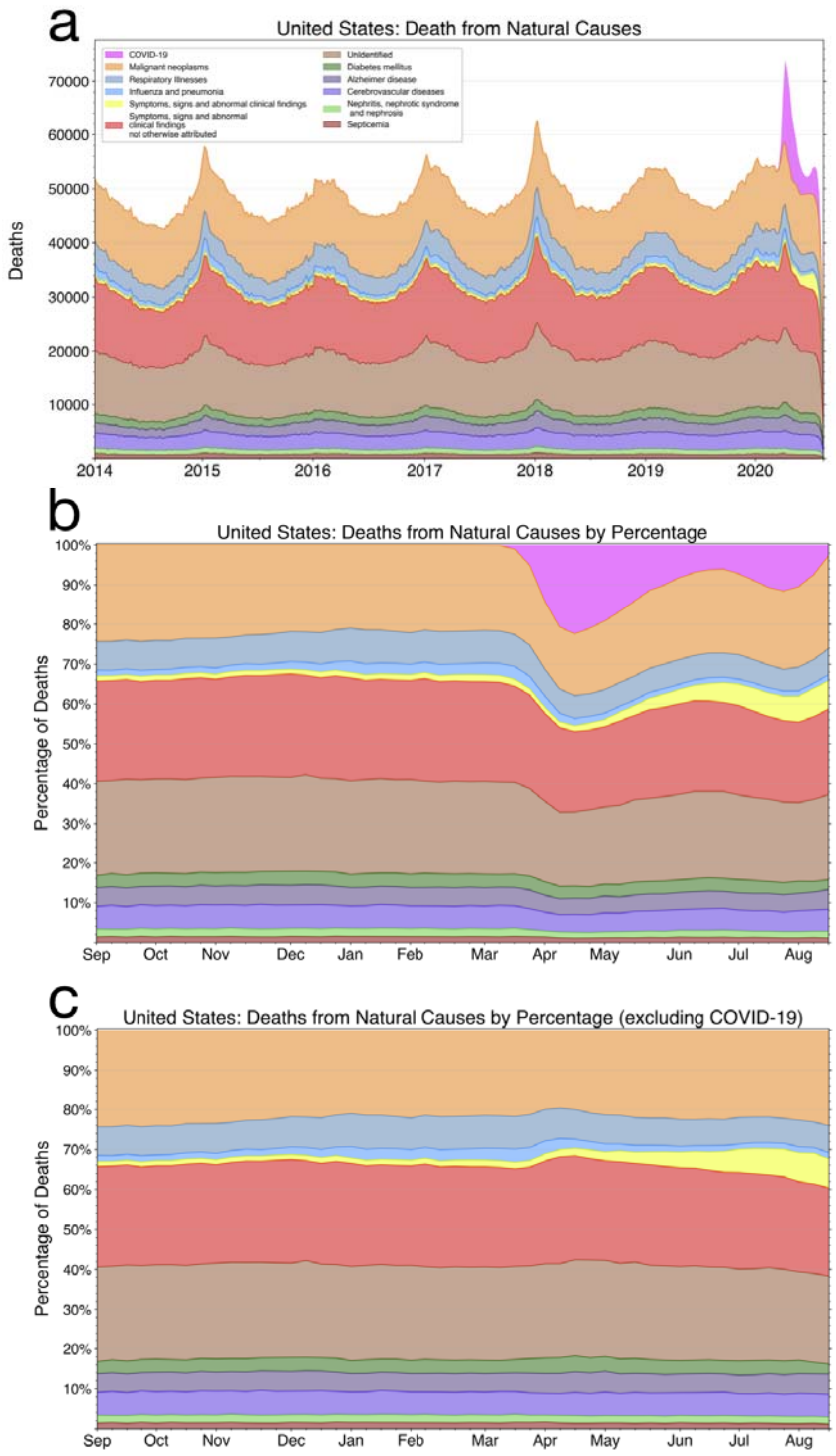


Figure 1. Numbers and percentages of deaths from Natural Causes in the United States

a) Number of deaths from 'Select Causes' from January 2014 - 16 August 2020

b) Percentages of deaths from 'Select Causes' including COVID-19, from 1 September 2019 - 16 August 2020

b) Percentages of deaths from 'Select Causes' excluding COVID-19, from 1 September 2019 - 16 August 2020

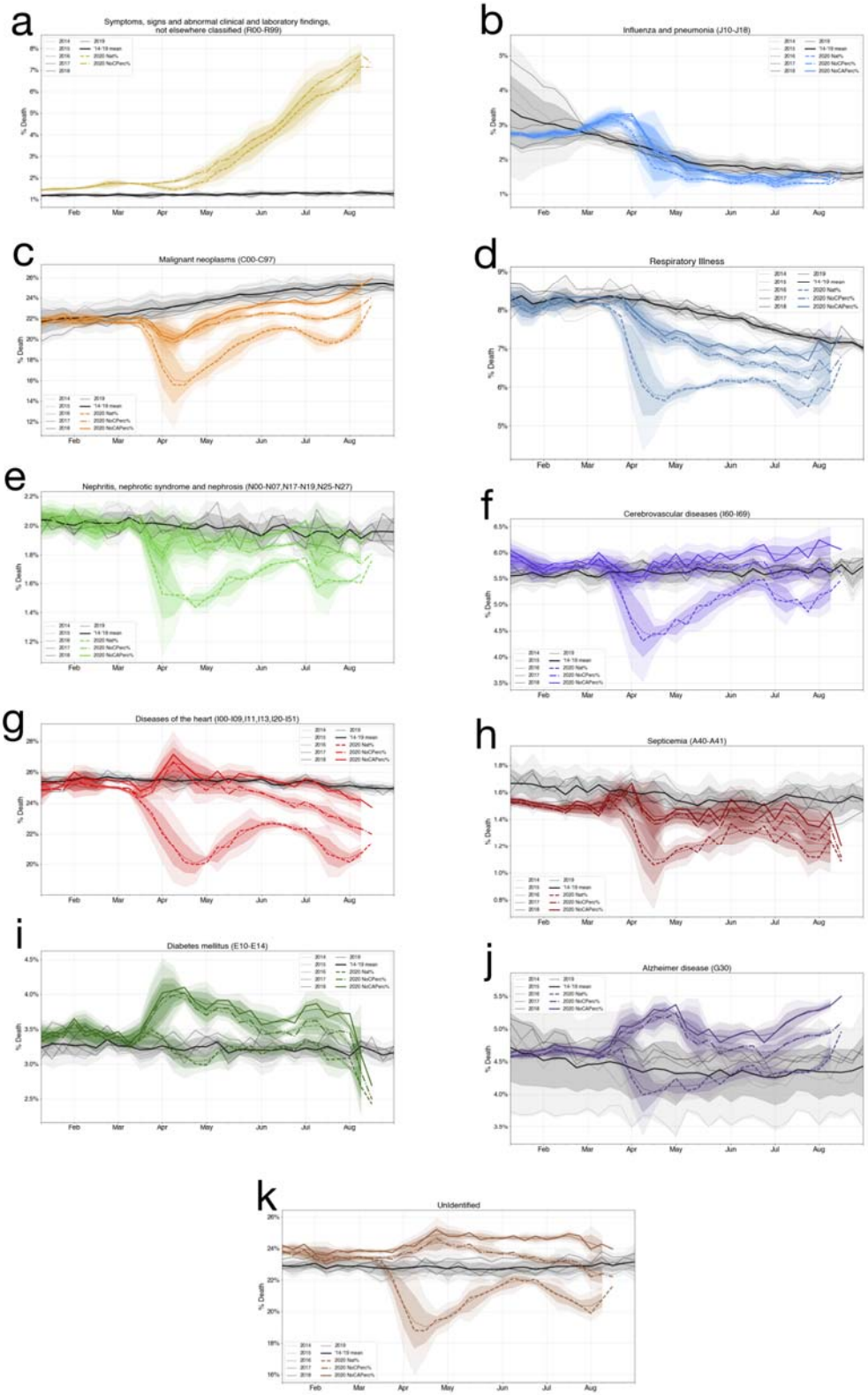


Figure 2. Percentages of US-wide deaths from Select Causes in 2020 compared to prior-year mean values.

Grayscale traces indicate mean and individual-year percentages from matched MMWR week in 2014-2019, as indicated. Colored traces indicate 2020 percentages calculated as a fraction of all Natural Cause deaths; Natural Cause deaths minus deaths from COVID-19 as an Underlying Cause; or Natural Cause deaths minus deaths from both COVID-19 as an Underlying Cause and Abnormal Findings, with thick lines indicating raw numbers and thin lines indicating the means of centered three-week running windows. Dark and light shading indicates one and two standard deviations from the respective mean values. Individual panels show deaths from a) Abnormal Findings; b) Influenza and Pneumonia; c) Cancer; d) Other Respiratory Illnesses; e) Nephritis; f) Stroke; g) Heart Disease; h) Septicemia; i) Diabetes; j) Alzheimer's; and k) Unidentified causes.

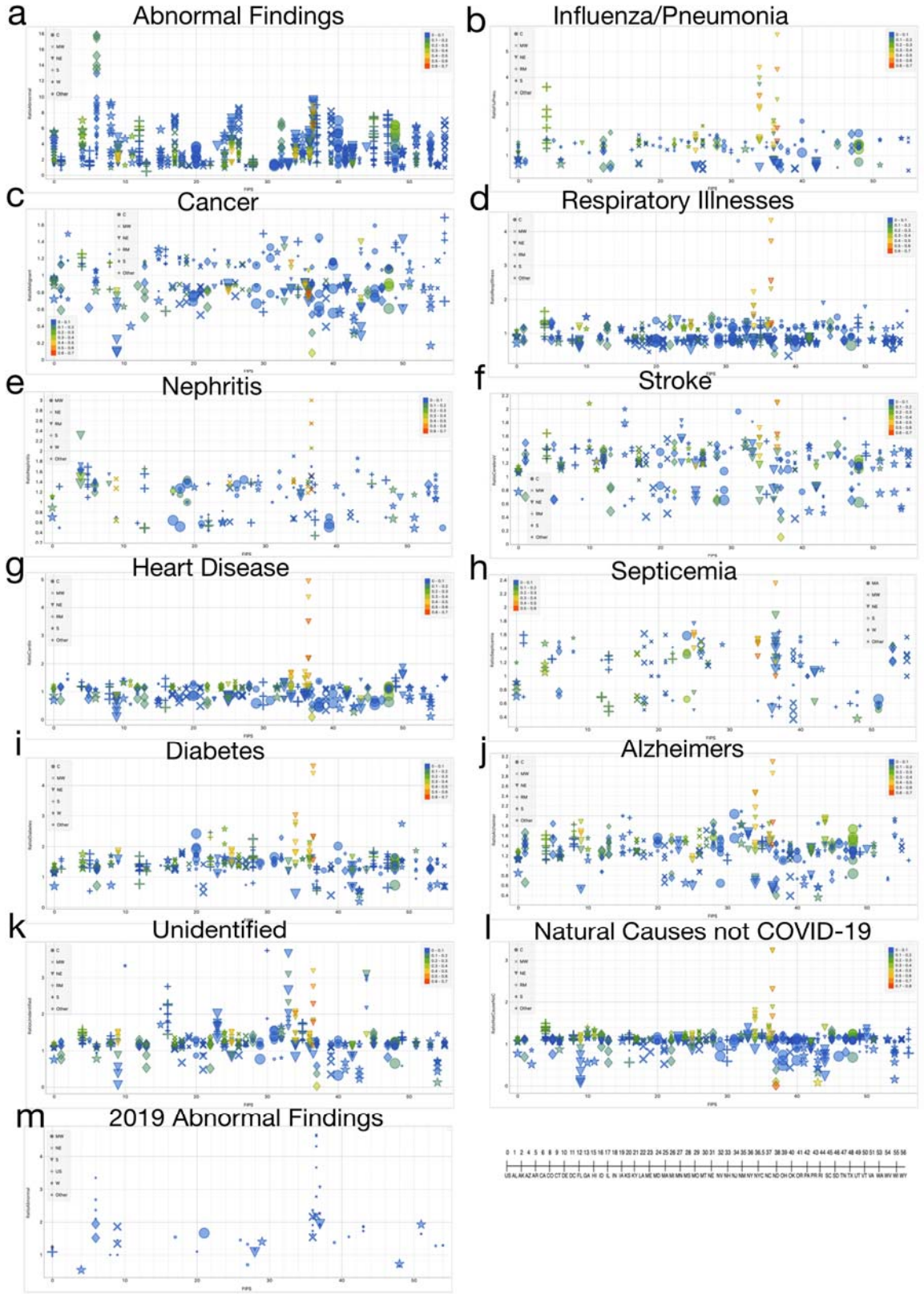
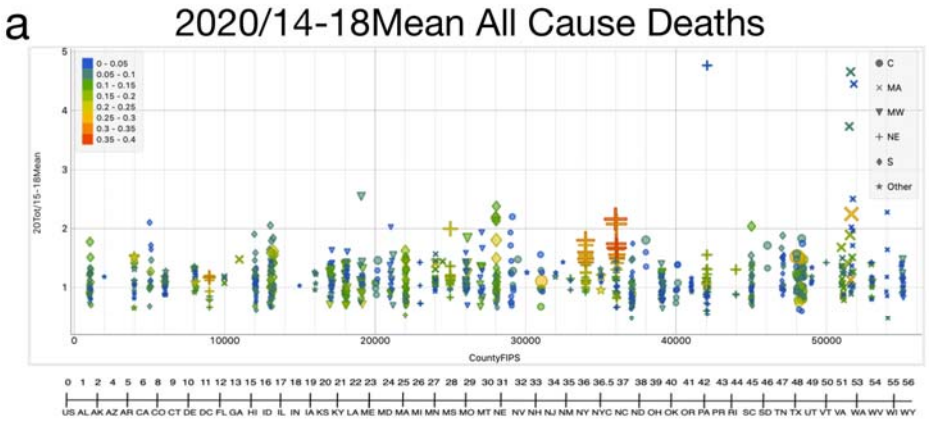
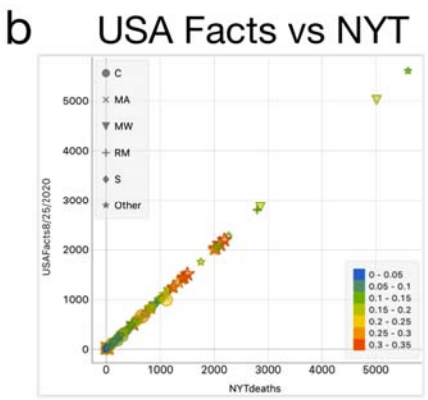
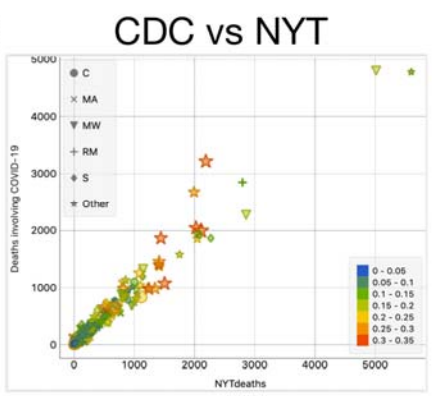
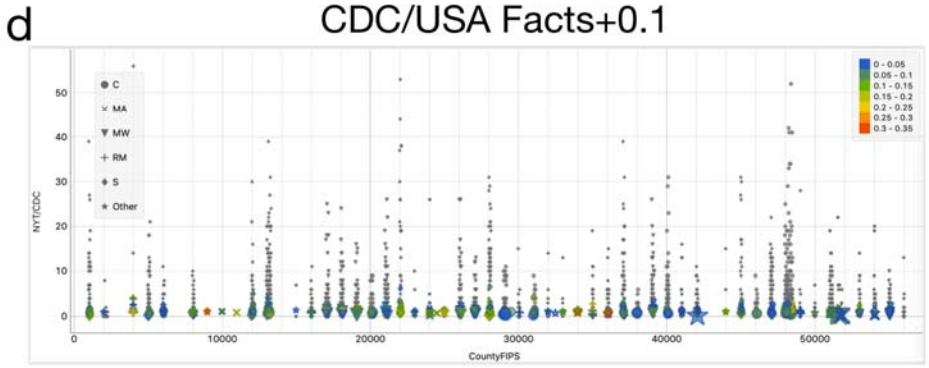


Figure 3. Ratios of deaths from Select Causes in 2020 divided by 2014-2019 mean values, by US state, the District of Columbia, New York City, and Puerto Rico, showing only values that differ significantly from prior-year means at 95% confidence. State FIPS code is shown on the X-axis, with corresponding state abbreviations shown at lower right. Point size corresponds to MMWR week, with larger size indicating more recent weeks, and point color indicates percent of deaths from COVID-19 as an Underlying Cause. Panels a-l show deaths between 1 February and 8 August 2020 from a) Abnormal Findings; b) Influenza and Pneumonia; c) Cancer; d) Other Respiratory Illnesses; e) Nephritis; f) Stroke; g) Heart Disease; h) Septicemia; i) Diabetes; j) Alzheimer's; k) Unidentified causes and l) Natural Causes excluding COVID-19. Panel m) shows deaths from Abnormal Findings between 15 December and 1 February that exceed mean 2014-2018 values with 95% confidence.









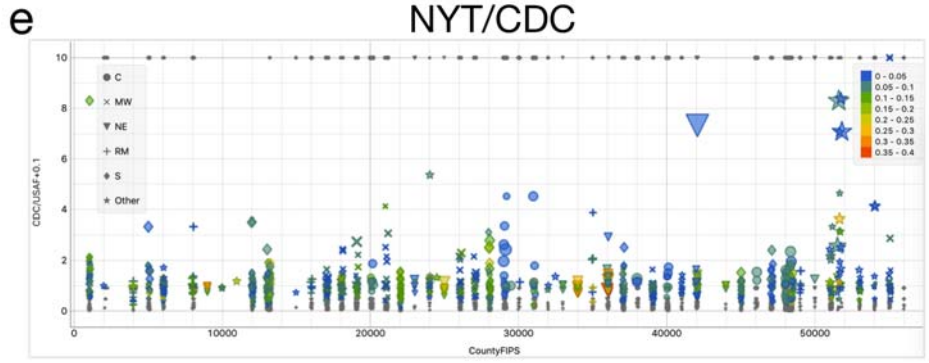


Figure 4. Cumulative deaths from COVID-19 between 1 February and 12 August, at county level. County FIPS codes are shown on the X-axis, corresponding to state FIPS code times 1,000, as indicated. Each point represents a single county, with color indicating the percent of deaths from COVID-19 as an Underlying Cause as indicated in the panel key, and point size corresponding to the ratio of deaths from COVID-19 as an Underlying Cause in 2020 to mean deaths from All Causes in 2014-2019 over the period 1 February and 12 August. Small grey points indicate counties that have reported no deaths from COVID-19 to the CDC, as of 26 August.

a) Ratio of deaths from All Causes in 2020 divided by the mean of deaths from All Causes in 2014-2019.

b) Cumulative number of deaths from COVID-19 as counted by the New York Times on the X-axis, and COVID-19 deaths as counted by USA Facts on the Y-axis.

c) Cumulative number of deaths from COVID-19 as counted by the New York Times on the X-axis, and COVID-19 deaths as recorded by the CDC on the Y-axis.

d) Ratio of cumulative COVID-19 deaths recorded by the CDC divided by COVID-19 deaths counted by USA Facts plus 0.1.

e) Ratio of cumulative COVID-19 deaths counted by the New York Times divided by COVID-19 deaths recorded by the CDC.

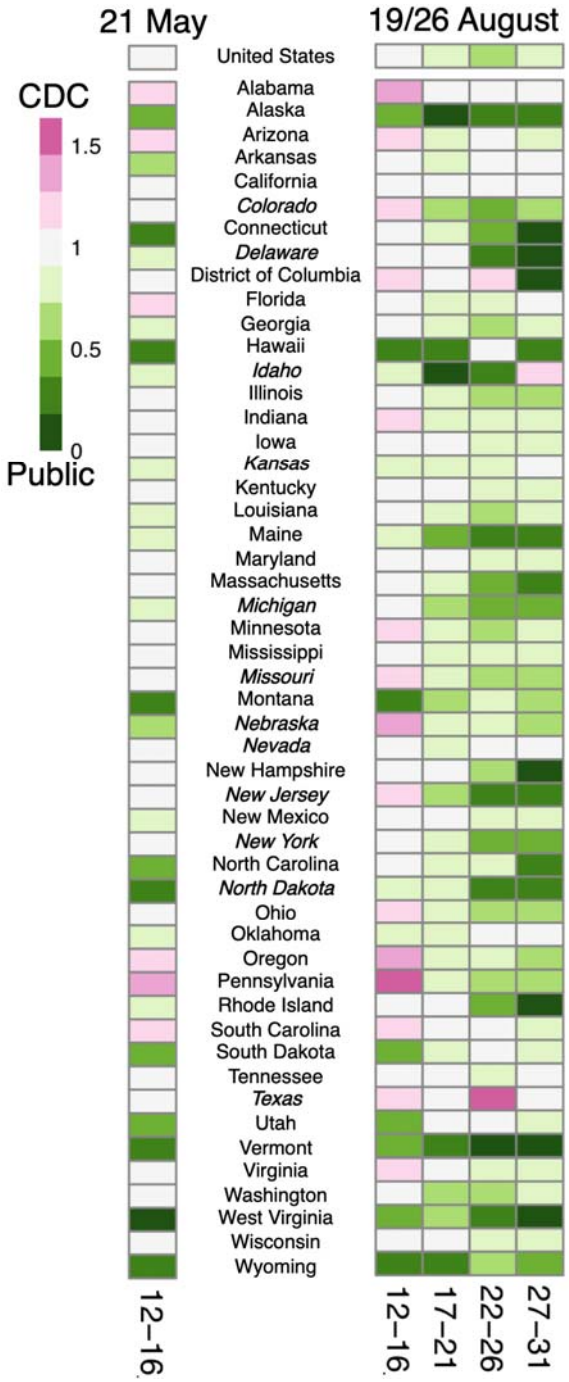


Figure 5. Heatmap indicating the ratio of COVID-19 deaths, binned by MMWR week, from datasets released by the CDC on 21 May or 26 August 2020, divided by the average of COVID-19 deaths counted by the four public data trackers as of 26 August, for the entire US and individual states. States with death counts as recorded by the public data trackers that have coefficients of variations more than twice that of the entire US for at least 8 weeks are indicated in bold italics.

Table 1. Estimates of unreported deaths, MMWR weeks 12-31.

Time Period	Natural Causes	Atlantic Difference	NYT Difference	JHU/CSSE Difference	USAFacts Difference	Average Difference
21 May MMWR12-16	310441	-52450	-39202	-1782	-48213	-35411.7
19/26 Aug. MMWR12-16	325359	-82746	-70336	-34977	-78720	-66694.7
19/26 Aug. MMWR17-21	309719	35942	44969	45240	49298	43862.5
19/26 Aug. MMWR22-26	264845	77185	121476	86881	119300	101210.5
19/26 Aug. MMWR27-31	263166	16748	36124	29447	28239	27639.5
19/26 Aug. MMWR17-31	837730	129876	202570	161568	196837	172712.5