PNAS RESEARCH ARTICLE | APPLIED MATHEMATICS
POPULATION BIOLOGY

COVID-19 cases and deaths in the United States follow Taylor's law for heavy-tailed distributions with infinite variance

Joel E. Cohen[a,b,c,d,1], Richard A. Davis[c], and Gennady Samorodnitsky[e]

Contributed by Joel E. Cohen; received May 28, 2022; accepted August 9, 2022; reviewed by John Nolan and Zhengjun Zhang

The spatial and temporal patterns of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) cases and COVID-19 deaths in the United States are poorly understood. We show that variations in the cumulative reported cases and deaths by county, state, and date exemplify Taylor's law of fluctuation scaling. Specifically, on day 1 of each month from April 2020 through June 2021, each state's variance (across its counties) of cases is nearly proportional to its squared mean of cases. COVID-19 deaths behave similarly. The lower 99% of counts of cases and deaths across all counties are approximately lognormally distributed. Unexpectedly, the largest 1% of counts are approximately Pareto distributed, with a tail index that implies a finite mean and an infinite variance. We explain why the counts across the entire distribution conform to Taylor's law with exponent two using models and mathematics. The finding of infinite variance has practical consequences. Local jurisdictions (counties, states, and countries) that are planning for prevention and care of largely unvaccinated populations should anticipate the rare but extremely high counts of cases and deaths that occur in distributions with infinite variance. Jurisdictions should prepare collaborative responses across boundaries, because extremely high local counts of cases and deaths may vary beyond the resources of any local jurisdiction.

COVID-19 | lognormal–Pareto distribution | Taylor's law | fluctuation scaling | variance function

Significance

Variations in the cumulative reported SARS-CoV-2 cases and COVID-19 deaths by US county, state, and date exemplify Taylor's law of fluctuation scaling, a widespread ecological and epidemiological pattern. Specifically, on day 1 of each month from April 2020 through June 2021, each state's variance (across its counties) of cases is nearly proportional to its squared mean of cases. COVID-19 deaths behave similarly. The largest 1% of counts are approximately Pareto distributed, with a finite mean and an infinite variance. Finding infinite variance has practical consequences. Local jurisdictions (counties, states, and countries) that plan for prevention and care of largely unvaccinated people should anticipate rare but extremely high counts of cases and deaths, by preparing collaborative responses across boundaries.

By 9 May 2022, 226 countries and territories reported more than 517 million confirmed cases of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and more than 6.27 million deaths. The United States reported more confirmed cases (83.6 million) and more deaths (1,025,000) than any other country (1).

We report that cumulative counts of cases and deaths by county in states of the United States are consistent with Taylor's power law of fluctuation scaling (henceforth TL), which will be explained below. The exponent of TL is not distinguishable from two, which means that the coefficient of variation of counts by county is not distinguishable from a constant across states on any given date. The counties with the lowest 99% of counts have a distribution of counts well approximated by the lognormal distribution, while the counties with the highest 1% of counts have a distribution of counts consistent with a Pareto distribution with an infinite variance. Simple statistical models based on the lognormal–Pareto distribution reproduce TL scaling with exponent equal to two. Mathematical results explain the observed and simulated TL scaling in multiple samples from heavy-tailed distributions. Planning for prevention and care in a largely unvaccinated population should anticipate the rare but extremely high counts of cases and deaths that occur in distributions with infinite variance.

We now describe the data and TL. We test TL, the lognormal model for the lowest 99% of counts, and the Pareto model for the extreme upper tail, and give evidence for an infinite variance of counts. We then review our empirical and theoretical findings. *SI Appendix* reports simulations and mathematical analyses of models that interpret the empirical findings. We state precisely and prove mathematically that multiple samples from heavy-tailed distributions obey TL with exponent two under certain conditions. *SI Appendix* also gives additional discussion.

1. Data

In the United States, *The New York Times* has tabulated cumulative cases and cumulative deaths at the end of each day since the first reported confirmed case on 20 January 2021 (2). Cumulative cases and cumulative deaths are reported according to their location in a primary subdivision of the United States (the 50 states, plus possessions, territories, and Washington, D.C., all referred to as "states" henceforth) and, within each state, by secondary subdivision (county, parish, borough, or other equivalents of county, all referred to as "counties" henceforth). From now on, "cases" refer to cumulative cases and "deaths"

Author contributions: J.E.C. designed research; J.E.C., R.A.D., and G.S. performed research; J.E.C., R.A.D., and G.S. contributed new reagents/analytic tools; J.E.C. analyzed data; and J.E.C., R.A.D., and G.S. wrote the paper.

Reviewers: J.N., American University; and Z.Z., University of Wisconsin–Madison.

Competing interest statement: J.N. was one of 11 co-PIs on a Multidisciplinary University Research Initiative grant from the US Army Research Office in 2012–2018 along with R.A.D. and G.S., coauthors of this paper. That grant ended at the end of 2018, approximately 42 mo ago, overlapping a few months with the 48-mo exclusion period.

See online for related content such as Commentaries.

[1]To whom correspondence may be addressed. Email: cohen@rockefeller.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2209234119/-/DCSupplemental.

Published September 12, 2022.

refer to cumulative deaths on a given date. We shall use "count" or "counts" to refer to either cases or deaths.

We downloaded the file us-counties.csv with 1,436,628 lines of data on 19 June 2021. All calculations use Matlab (version 2021a). We select the first day of each of the 15 mo from April 2020 through June 2021 (keeping 47,004 lines of data). We exclude county–days with reported counts of zero (keeping 46,956 lines of data for positive cases and 37,649 lines for positive deaths), and we sort the counties by state. There were 843 month-state combinations, an average of $56.2 = 843/15$ states each month for 15 mo. We then exclude month–state combinations with six or fewer counties, leaving 718 month–state combinations with an average of $47.9 = 718/15$ states each month for 15 mo. We then compute the means and the variances of counts over the counties within each state on each of the 15 dates. Because each remaining state has seven or more counties, no remaining month–state combination has zero mean or zero variance over the counties within a state.

We attempt no adjustments for possible underreporting of cases or deaths due to COVID-19. If, as seems likely, different jurisdictions had different propensities to test, systematic data that would make it possible to adjust for such differences are not available. High counts of cases could either overwhelm local testing capacity, resulting in undercounts of tested cases, or could elicit supplemental test resources, resulting in unusually high reported cases. Such possible limitations of the data seem unlikely to have generated the systematic patterns we shall describe.

2. Taylor's Law (TL)

On each first day of 15 mo, the (mean, variance) pairs, one point for each retained state on that date, closely approximate TL, for both cases and deaths (Fig. 1). Specifically, on each date, each state's sample variance of the count (over its counties) is approximately proportional to some power of that state's sample mean of the count (over its counties). Equivalently, the logarithm of a state's sample variance is approximately a linear function of the logarithm of that state's sample mean. On log–log coordinates, there is no visual indication of systematic curvature. The ranges of the axes of all panels are the same within each figure, to make it easy see that, as time passes and the counts increase, the cloud of observed (mean, variance) pairs shifts from the lower left corner to the upper right corner.

In statistical language, TL says that the counts have a power variance function (3–5). Explicitly, for real constants $k > 0$ and b, both independent of the state i, but depending on the date and whether the counts are cases or deaths, TL proposes, and we find empirically, that

$$\text{sample variance of state } i \approx k \times (\text{sample mean of state } i)^b, \quad i = 1, 2, \ldots. \qquad [1]$$

Let $\log = \log_{10}$ here and throughout the data analysis, and let $a = \log_{10} k$. (In *SI Appendix*, the simulations use $\log_{10}$, and the mathematical theorems and proofs use the natural logarithm.) Then the power-law form of TL in Eq. **1** is equivalent to the linear log–log form of TL displayed in Fig. 1,

$$\log(\text{sample variance of state } i) \approx a + b \times \log(\text{sample mean of state } i), \; i = 1, 2, \ldots \qquad [2]$$

or the ratio form of TL,

$$\text{sample variance of state } i/(\text{sample mean of state } i)^b \approx k, \quad i = 1, 2, \ldots. \qquad [3]$$

The exponent b in Eqs. **1** and **3** is the same as the slope b in Eq. **2**, so b may be called either the slope or the exponent of TL. These specifications of TL intentionally leave vague the error model behind the approximation $\approx$.

TL is an empirical regularity widely observed in many sciences, including ecology, infectious disease epidemiology, human demography, financial statistics, earth sciences, and other physical sciences (6, 7).

The Poisson distribution, a common model of purely random variation in counts, has a variance equal to its mean. As the mean of a Poisson distribution increases to larger values, the Poisson distribution increasingly approximates a normal distribution with variance equal to the mean. Both the Poisson distribution and its normal approximation with variance equal to the mean follow TL with $k = b = 1, a = 0$. If states had different average counts per county and a Poisson distribution of counts over the counties within each state, then the states' means and variances of the counts would approximate TL with $k = b = 1, a = 0$. Graphically, on (log mean, log variance) coordinates, a family of Poisson distributions with varying mean will lie on or near a line of slope one through the origin. Fig. 1 shows that the Poisson distribution does not describe even approximately how the sample variance of counts by county within states relates to the sample mean. We infer that other sources of variation besides purely random fluctuation influence the counts, such as heterogeneity or contagion.

Only once in Fig. 1 does the 95% CI of the slope exclude $b = 2$. For deaths in April 2020, the estimated slope is 3.003 with 95% CI (2.685, 3.321). In all the remaining 29 instances of TL ($29 =$ 15 mo $\times$ 2 counts (cases or deaths) -1), the estimated slope b of TL in Eq. **2** is statistically indistinguishable from two. When $b = 2$, the ratio (sample variance)/(sample mean)^2 is independent of any positive rescaling of the original measurements.

In every month, for both cases and deaths, the lower bound of the 95% CI of the slope b exceeds one, excluding Poisson variation among counties in cases and deaths.

The parameters a, b of the log–log form of TL Eq. **2** (Fig. 2) show no substantial trends over time. For both counts, early in the period of observation, the intercept a rises slightly while the slope b falls slightly, as if the fitted straight line that represents TL were rotating slightly clockwise. In the second half of the period of observation, there is no suggestion of change in the TL parameters.

3. Lognormal Model of TL with Slope Two

To understand why TL holds and why the slope b in Eq. **2** approximates two (except for deaths in April 2020), we observe that the spread of infection could be modeled by a variety of stochastic multiplicative processes. For example, the supercritical discrete generation Galton–Watson branching process with finite mean > 1 and finite variance > 0 of the offspring distribution satisfies TL with slope two asymptotically in time (ref. 8, p. 33). The supercritical continuous-time birth and death process with a birth rate per individual that strictly exceeds the death rate per individual also satisfies TL with slope two asymptotically in time (ref. 8, pp. 33–34). These and other similar examples provide prototypes of explanations of why TL holds with slope two in the COVID-19 counts. But the details of transmission of infection and death from COVID-19 are not adequately described by these simple models.

A more robust explanation is required that does not depend on the details of transmission. The lognormal and Weibull distributions are limiting distributions of large families of mechanisms,

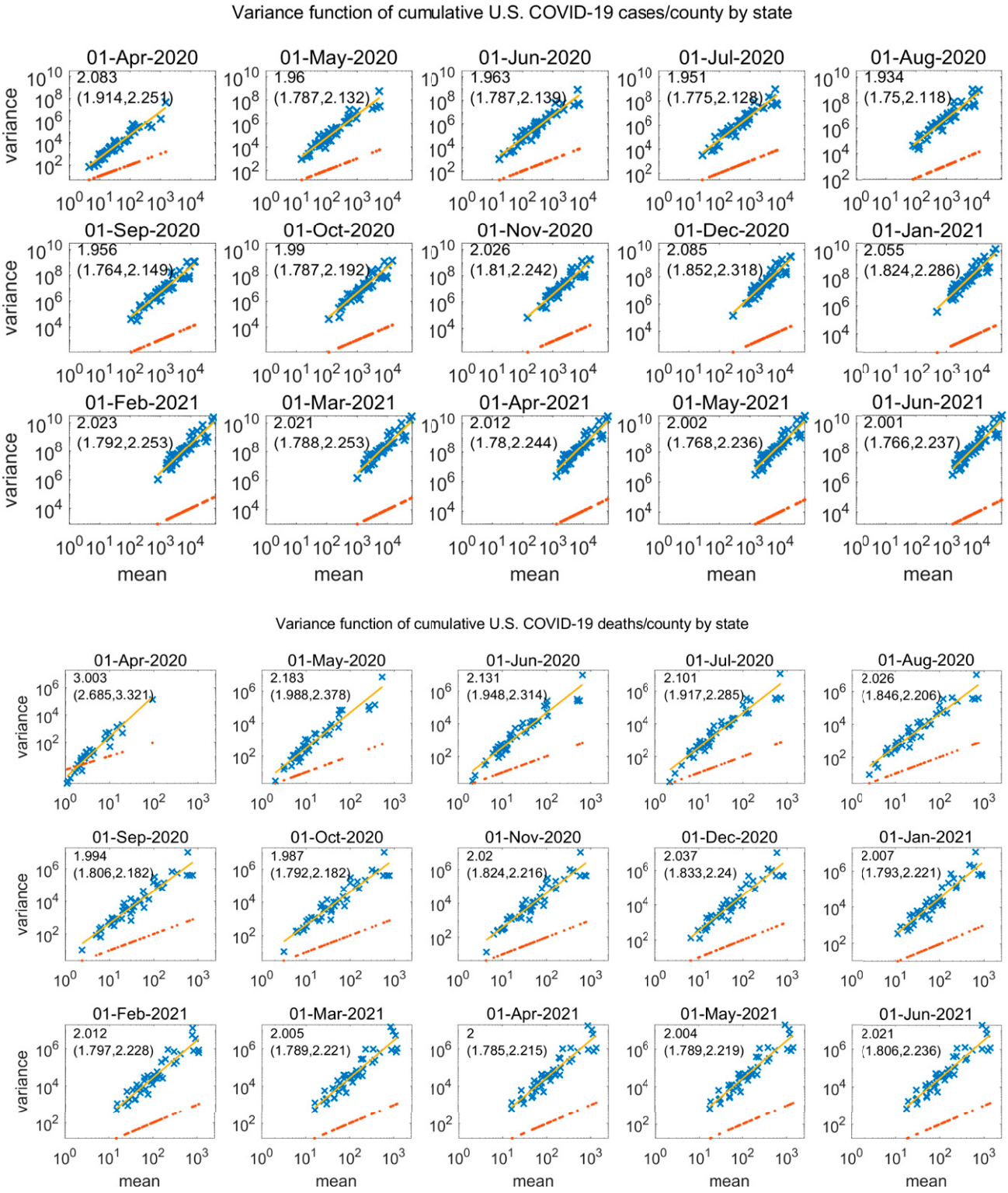

Fig. 1. Cumulative US COVID-19 cases (*Upper*) and deaths (*Lower*) by state are well described by TL. Each blue × marker shows, on log–log coordinates, the (mean, variance) of the number of cases across counties within one state. The yellow straight line fitted by ordinary least squares to the blue × markers (log mean, log variance) is the estimated TL. In all 15 mo, R^2 is between 0.90 and 0.92 for cases and between 0.87 and 0.91 for deaths. The estimated slope b of TL is the top left figure in each panel. An approximate 95% CI of the slope is given in parentheses below the estimated slope b. As the months pass, the estimates of b become increasingly and remarkably close to two. The red dots in a straight line near the bottom of each panel show a hypothetical variance equal to the mean, as in a family of Poisson distributions with parameters equal to the mean count of each state. The similarity in the bottom five panels reflects the temporary slowdown in new COVID-19 cases in the first half of 2021.

thanks to the central limit theorem (9). For example, Yule-type multiplicative growth and division processes converge asymptotically in time to the lognormal distribution and the Weibull distribution under different conditions (10).

So far, we have analyzed the distribution of counts over counties within each state separately. Because any single state has few counties for the purpose of discriminating between lognormal and Weibull distributions or of examining the upper tail of the

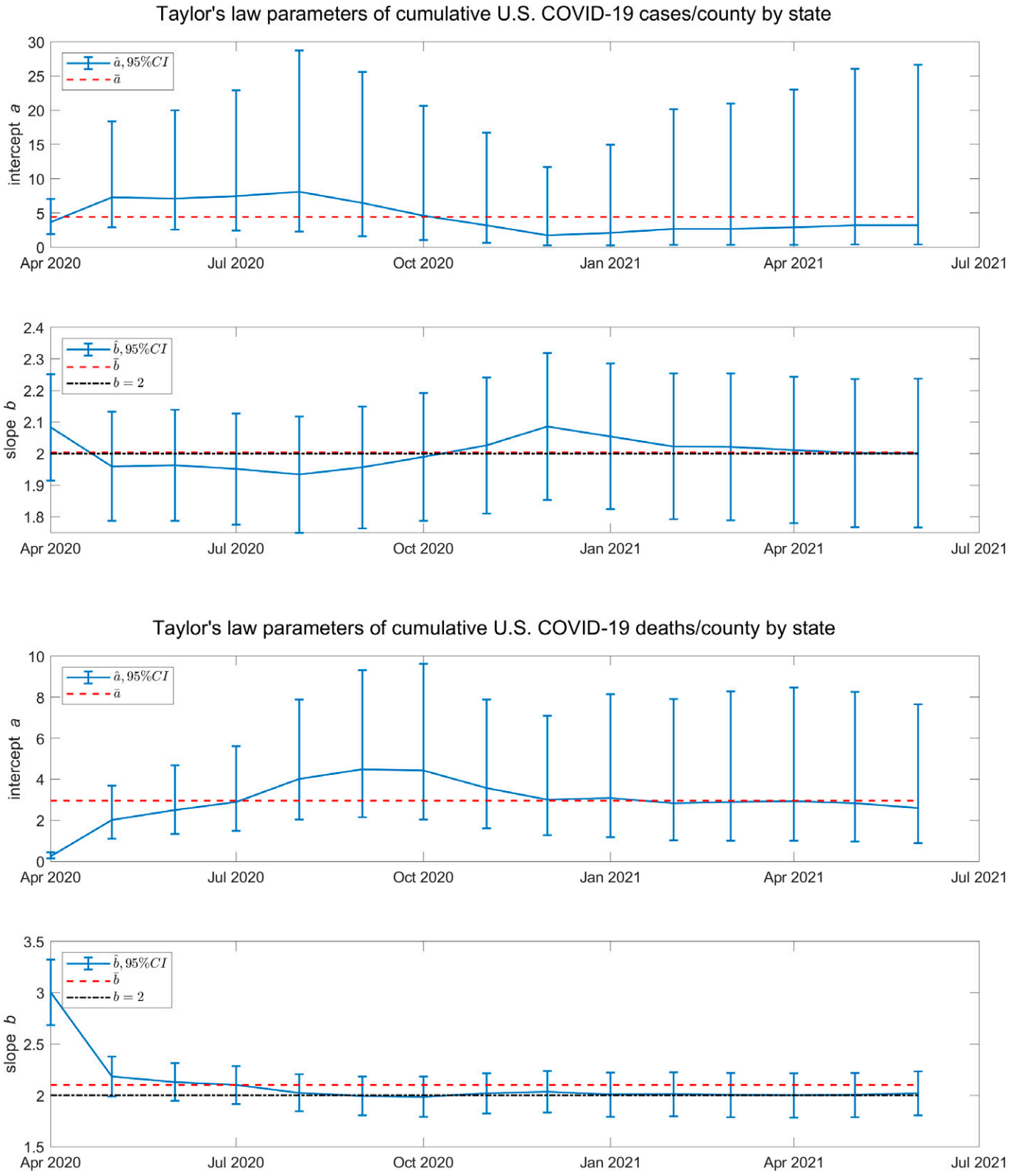

Fig. 2. Intercept *a* (panels 1 and 3) and slope *b* (panels 2 and 4) of TL Eq. **2** fitted to cumulative US COVID-19 cases (panels 1 and 2) and deaths (panels 3 and 4). The solid blue line joins the point estimate of the parameter on each date. The upper and lower error bars show the 95% CI. The horizontal dashed red line shows the average value of the point estimates of the parameter over the period of observation. In panels 2 and 4 for the slope, the horizontal dash-dotted black line shows $b = 2$. This line falls within the CI in every month (except for deaths in April 2020) and nearly coincides with the dashed red line for $\bar{b}$.

survival curve, we now (temporarily) analyze the distribution of counts over all counties within the United States. In this analysis, all counties with a positive count are considered, regardless of the state in which they occur.

Fig. 3 plots, on log–log coordinates, the empirical survival curve $\Pr(X > x)$ of the count X as a function of the positive number x for all US counties with positive count on day 1 of each of 15 mo, along with the survival curves of lognormal and Weibull

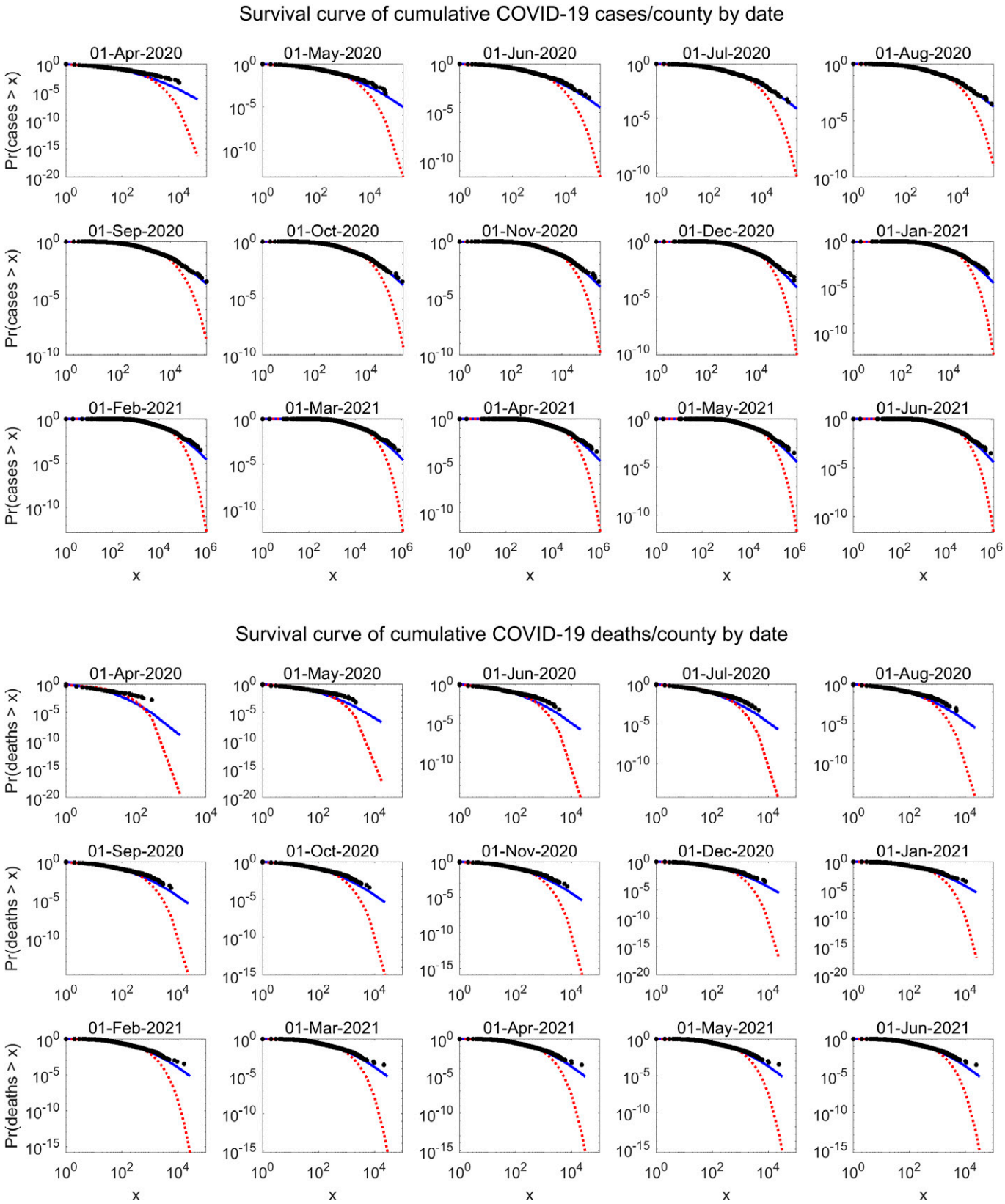

Fig. 3. Empirical survival curve $\Pr(X > x)$ as a function of x = cumulative COVID-19 cases (*Upper*) and deaths (*Lower*), for all US counties included in the previous analyses, on log–log coordinates. The solid blue curve is the lognormal distribution fitted by maximum likelihood to all counties, including those in the upper tail. The dotted red curve is the Weibull distribution fitted by maximum likelihood to all counties, including those in the upper tail. Each black dot $(x, P(X > x))$ represents the count x in one U.S. county and the fraction $P(X > x)$ of all counties with a count larger than x.

distributions fitted by maximum likelihood to the counts of all the counties included in each panel. If cases were Pareto distributed, the empirical survival curve would be a downward sloping straight line on log–log coordinates. The plots are clearly concave on log–log coordinates, not linear. The curvature appears to increase as the months pass. Both the lognormal and the Weibull survival

curves describe roughly 99% or more of the empirical survival curve (corresponding to the interval from 10^0 at the top of the vertical axis down to 10^{-2}). However, the Weibull survival curves fall off much faster than both the empirical survival curves and the lognormal survival curves. Henceforth, we disregard the Weibull distribution as a model for these data.

The lognormal distribution, while far better than the Weibull, is imperfect: For the largest values of the threshold x, the lognormal survival curve always falls off faster than the empirical survival curve. This discrepancy means that extremely large counts have higher probability in the data than predicted by the best-fitting lognormal distribution. Extremely high counts are more likely than the lognormal model would predict, despite the good agreement with the lognormal model for 99% of counts. In the top percentile upper tail where the fitted lognormal drops below the data, the empirical survival curve appears to fall roughly linearly on log–log coordinates, like a Pareto distribution. We examine the top percentile upper tail of the empirical survival curve in Section 4.

A positive-valued random variable $Y(\mu, \sigma^2)$ with real parameters μ and $\sigma^2 \geq 0$ is lognormal if $\log(Y(\mu, \sigma^2))$ is normal with mean μ and variance σ^2. The mean and the variance of $Y(\mu, \sigma^2)$ are

$$E[Y(\mu, \sigma^2)] = \exp(\mu + \sigma^2/2), \quad [4]$$

$$var[Y(\mu, \sigma^2)] = [\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2) = [\exp(\sigma^2) - 1]\{E[Y(\mu, \sigma^2)]\}^2. \quad [5]$$

This variance function Eq. **5** of the lognormal distribution can exhibit various behaviors, depending on the relation between σ and μ (Appendix in ref. 11). If each state's distribution of counts is well approximated by a lognormal distribution, and if the parameter μ of $Y(\mu, \sigma^2)$ changes while σ^2 remains constant (or varies little, compared to variation in μ) from state to state on a given date, then the coefficient $[\exp(\sigma^2) - 1]$ on the right side of Eq. **5** remains constant (or varies little) while the mean $E[Y(\mu, \sigma^2)]$ changes. Then the variance changes in proportion to the square of the mean. Thus, if σ^2 is constant (or varies little) and if μ varies from state to state on a given date by amounts that are large compared to the variation in σ^2, this variance function Eq. **5** obeys TL with $b = 2$.

To see whether the counts, state by state, support this proposed explanation of our empirical finding in Fig. 1 that TL holds with $b \approx 2$ in 29 of 30 cases, we return to data analysis at the state level. We fit lognormal distributions by maximum likelihood to all the positive counts per county of each state on each date separately. Fig. 4 plots, for each date, the state-specific estimates of the parameter μ on the abscissa and, on the ordinate, σ (the square root of the parameter σ^2, to get both axes on the same scale).

For cases (Fig. 4, *Upper*), the slope of a straight line fitted by ordinary least squares to the points (μ, σ), one point per state, does not differ significantly from zero in every month after April 2020. The cloud of points moves to the right from $\mu \in (0, 6)$ in April 2020 to $\mu \in (6, 11)$ in June 2021, while $\sigma \in (0, 3)$ over the entire period. At each date, and over all months, μ ranges much more widely than σ.

For deaths (Fig. 4, *Lower*), the slope of a straight line fitted by ordinary least squares to the points (μ, σ), one point per state, declines steadily from April 2020 through November 2020. From October 2020 through June 2021, the 95% CI of the slope includes zero. The cloud of points moves to the right from $\mu \in (0, 3)$ in April 2020 to $\mu \in (1, 8)$ in June 2021, while $\sigma \in (0, 3)$ over the entire period. Again, at each date, and over all months, μ ranges more widely than σ. The slope of σ as a linear function of μ exceeds zero only early in the pandemic.

To a first approximation, the model of a lognormal distribution with fixed parameter σ^2 and changing parameter μ describes the empirical lognormal parameter estimates well for cases in all months except April 2020, and well for deaths from October 2020 onward, thereby explaining why, in most cases, TL holds with $b \approx 2$.

This explanation is incomplete in at least three respects. First, the model leaves open the question of why the lognormal parameters behave as they do. Second, the model does not interpret why, in the early months of the epidemic, σ is larger in states with larger μ. For cases, the increase of σ with μ is limited to April 2020. For deaths, the increase of σ with μ is more dramatic and extends over several early months. Third, the model does not interpret why, for both counts in every month, the extreme upper tail of the empirical survival curve of all counties in the United States falls like a Pareto distribution, more slowly than the lognormal distribution.

The observation that, in the early months of the epidemic, σ is larger in states with larger μ explains why TL slopes are larger than two in the early months. When σ is larger in states with larger μ, it is obvious from the variance function Eq. **5** that the variance will increase faster than the square of the expectation, and a fitted TL will have an estimated slope larger than two. For example (Appendix in ref. 11), if $\sigma^2 = \mu$, then TL holds approximately, and its slope is $2 + 2/3$, while, if μ is constant and only σ^2 varies, then TL holds approximately with slope four, as for tornadoes (*SI Appendix*). These examples provide a qualitative insight into why, in the early months, TL slopes larger than two are associated with σ being larger in states with larger μ (for cases, April 2020; for deaths, April 2020 through roughly October 2020).

4. Lognormal–Pareto Model for the Survival Curves of Counts

The survival curves of counts in all counties of the United States (Fig. 3) are consistently well described by a lognormal distribution for the lower 99% or more of values of the count. For the highest ∼1% of counts, the empirical survival curves all fall more slowly than the fitted lognormal survival curves. On log–log coordinates, the empirical survival curves fall approximately linearly. This combination of lognormal and Pareto behavior was known for decades from graphical examples (ref. 12, pp. 110–111 and figures 1 and 2; ref. 13, p. 219 and figure 1; ref. 14, pp. 3, 5, and 6 and figures 1–3) before it was formalized and named the lognormal–Pareto distribution (15–17). The lognormal–Pareto distribution (in its version with a continuous and differentiable probability density function [pdf]) is specified by a threshold $\theta > 0$, a tail index $\alpha > 0$, and a scatter $\sigma > 0$. For $x > 0$, its pdf $f(x) := rf_1(x; \mu, \sigma^2, \theta) + (1 - r)f_2(x; \theta, \alpha)$, $r \in (0, 1)$ is a weighted sum of a lognormal distribution right-truncated at θ with pdf

$$f_1(x; \mu, \sigma^2, \theta) := \left[\Phi((\log(\theta) - \mu)/\sigma)x\sigma\sqrt{2\pi}\right]^{-1} \exp\{-((\log(\theta) - \mu)/\sigma)^2/2\}\mathbf{1}_{\{0<x\leq\theta\}}$$

and a Pareto distribution with left threshold θ and pdf $f_2(x; \theta, \alpha) := \alpha\theta^\alpha x^{-(\alpha+1)}\mathbf{1}_{\{\theta<x\}}$. Because $f(x)$ is required to be continuous and differentiable in x, parameters r and μ are functions of the other parameters (16, 17).

To examine in greater detail the upper tail of the survival curves in Fig. 3, we plot $\log \Pr(X > x)$ as a function of $\log x$ for only the counties with the highest 1% of cases (Fig. 5, *Upper*) or deaths (Fig. 5, *Lower*). The number of such counties ranges from a low of

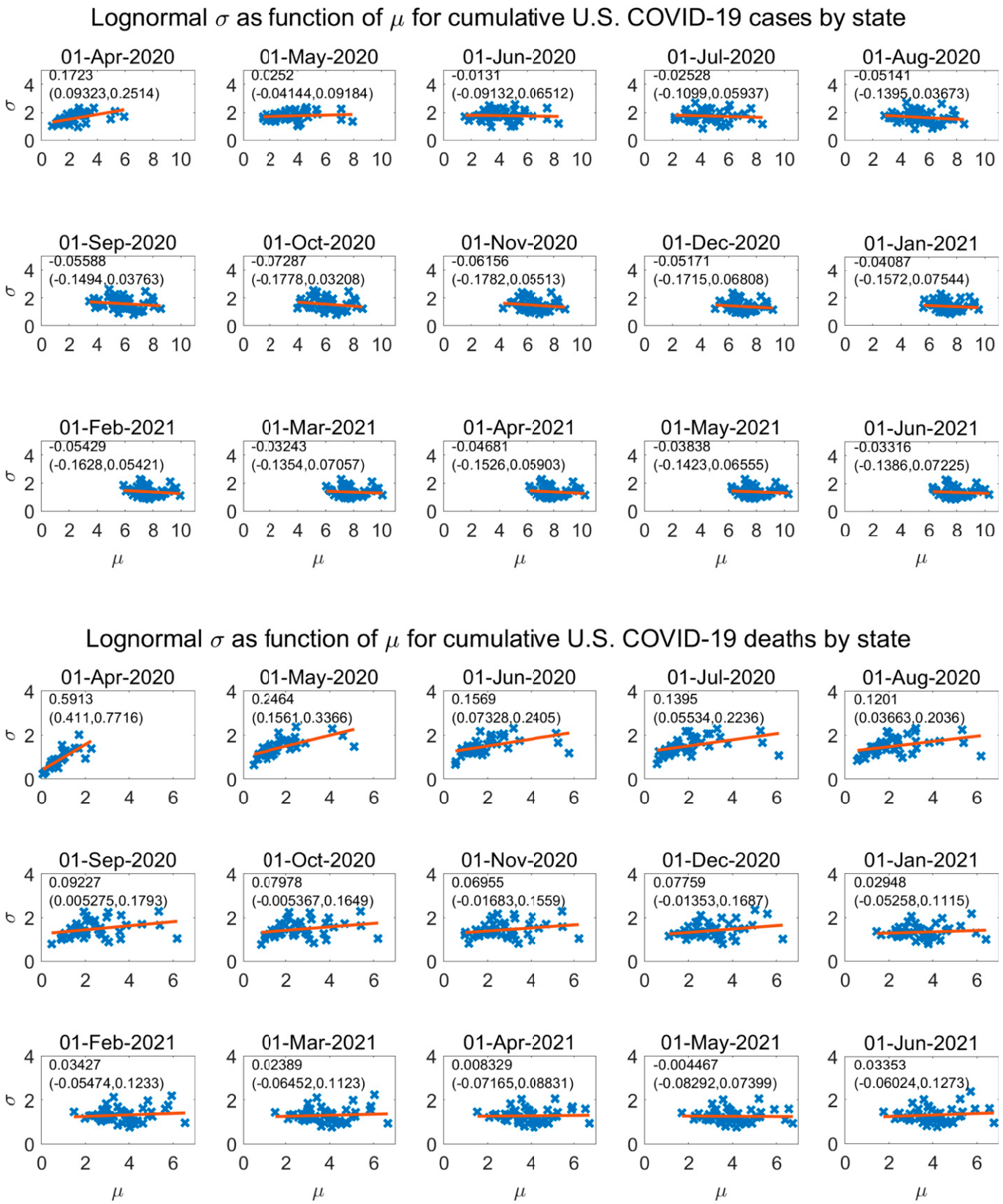

Fig. 4. Lognormal distributions are fitted by maximum likelihood to the cases (*Upper*) or deaths (*Lower*) per county within each state separately, that is, fitted to $x_{1j}, \ldots, x_{r_j j}$ separately for each state $j = 1, \ldots, c$, where r_j is the number of counties in state j. Each panel plots the lognormal parameter σ_j of each state j as a function of the lognormal μ_j (blue $\times$ markers). To a first approximation, μ_j varies over a considerably wider range than does σ_j. A solid red straight line is fitted using ordinary least squares. The slope of the fitted line is given in the upper left corner of each panel. The 95% CI, given below the slope, includes zero for all months except April 2020.

six (once only, for deaths in April 2020, before COVID-19 deaths had spread widely) to 32 (Fig. 6). On each date, we consider the counts x (of cases or deaths) of all counties regardless of state.

In each panel of Fig. 5, the horizontal axis is labeled, on the left, with the lowest count included in the largest 1% of counts and, on the right, with the highest count. On visual inspection,

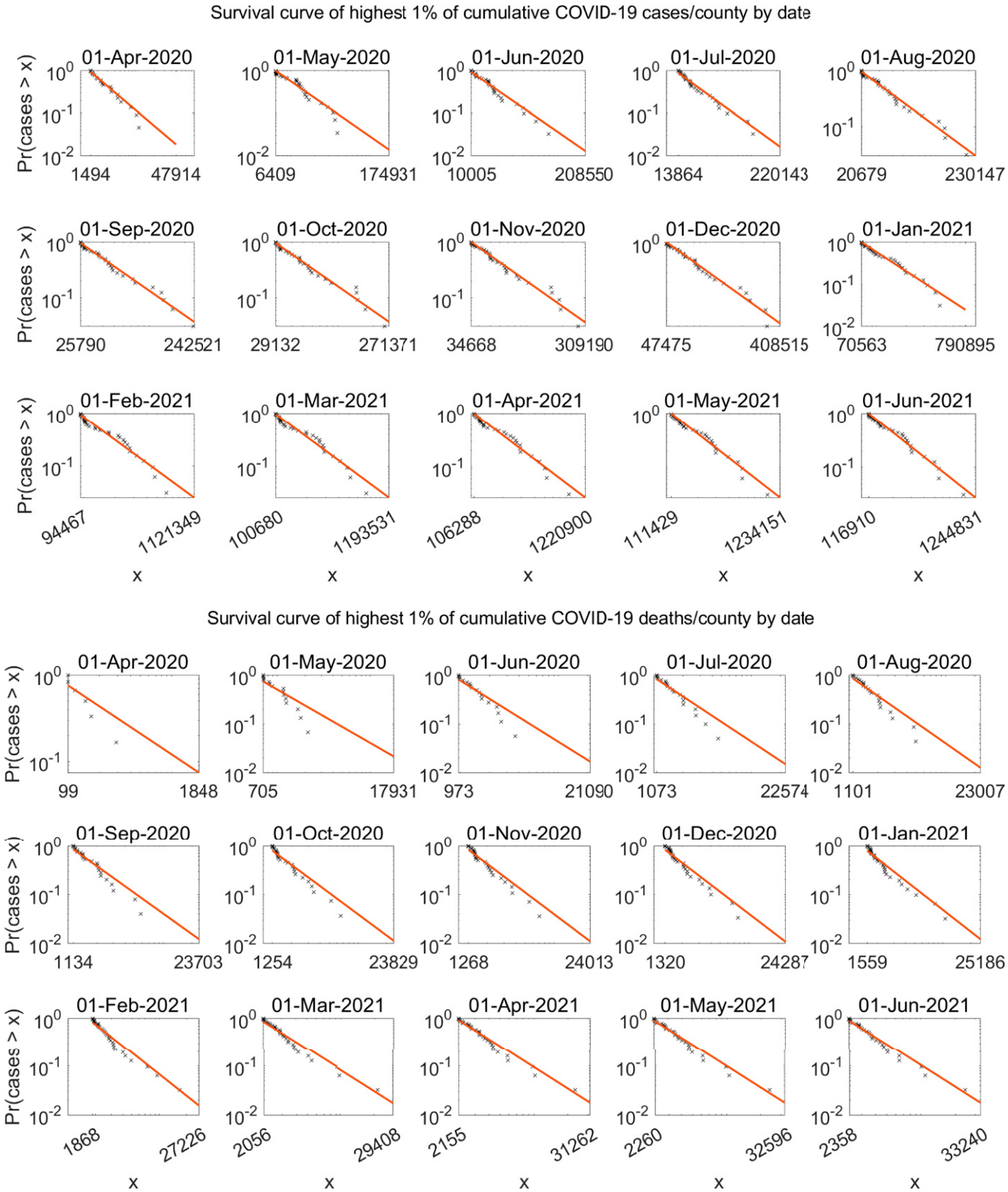

Fig. 5. Top percentile of the empirical survival curve $\Pr(X > x)$ as a function of the cumulative count x of COVID-19 cases (*Upper*) or COVID-19 deaths (*Lower*) of one county on that date, on log–log coordinates. Each county is represented by one blue $\times$. The orange straight line is fitted by the method of ref. 21.

a linear relationship of $\log \Pr(X > x)$ to $\log x$ seems reasonable, as specified in a Pareto distribution. If this linear relationship is written $\log \Pr(X > x) = \beta - \alpha \log x$, α is called the tail index (or simply the index) of the Pareto distribution. The value of the tail index determines whether the Pareto distribution (and hence also the lognormal–Pareto distribution) has finite or infinite mean (in case $\alpha > 1$ or $\alpha \in (0, 1]$, respectively) and, if the mean is finite, whether the variance is finite or infinite (in case $\alpha > 2$ or $\alpha \in (1, 2)$, respectively).

For the counties with the largest 1% of counts, we estimate α using the Hill estimator (ref. 18 and ref. 19, p. 69, section 3.2.2). Let $X_{(1)} \geq \ldots \geq X_{(m+1)}$ be the $m + 1$ largest order statistics of

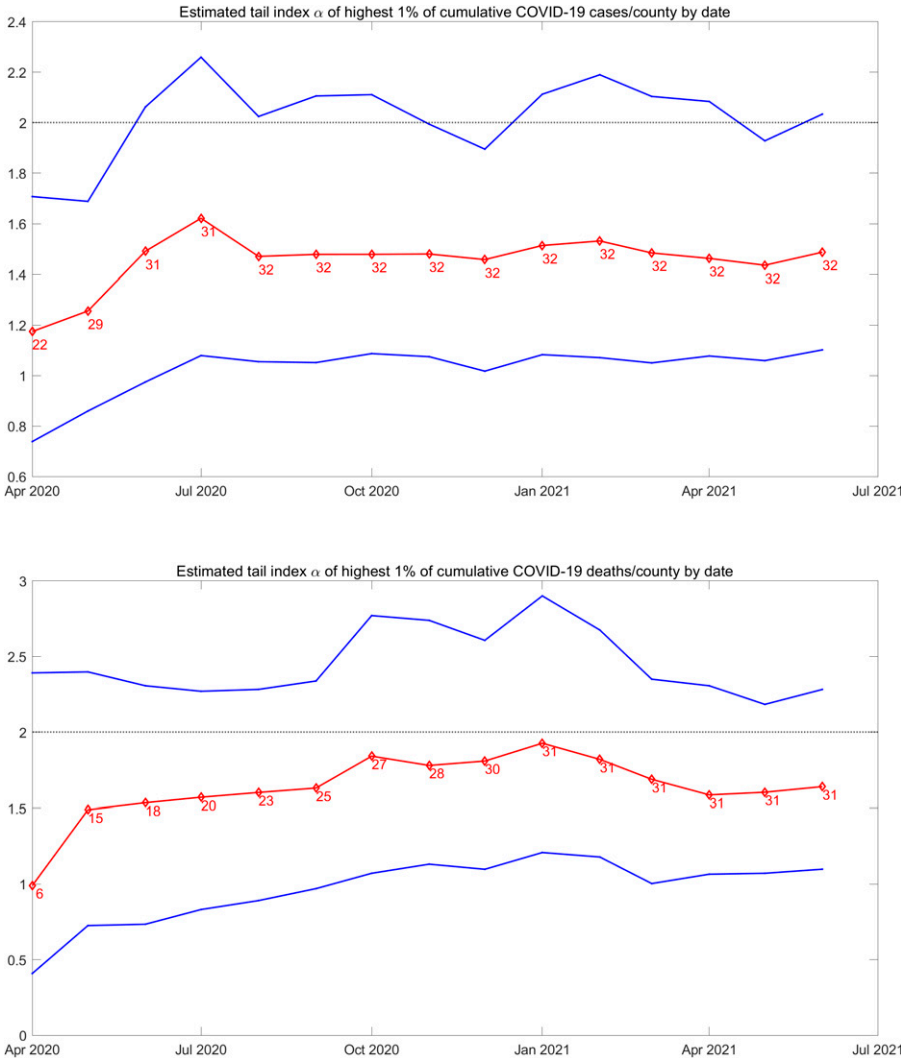

Fig. 6. For each month (horizontal axis), Hill estimates (ref. 18 and ref. 19, p. 69, section 3.2.2) (vertical axis) of the tail index α (red diamond markers, red solid line) of the empirical survival curve $\Pr(X > x)$ for the counties with the largest 1% of cases (*Upper*) and deaths (*Lower*), and 95% CIs (solid blue lines) of the point estimates of α based on 1,000 bootstrap samples with replacement of the counties with the largest 1% of counts. The number with each data point is the number of counties in the top 1% of counties on that date; it corresponds to $m + 1$ in Eq. **6**.

all counts $X_k,\ k = 1, \ldots, n$ of the n counties on a given date. Since the Pareto model describes only the largest counts, inference about α will be based on only the largest 1% of counts. The maximum likelihood estimator of α based on the largest 1% of counts on a given date is (using the natural logarithm here)

$$\hat{\alpha} = \left[\frac{1}{m} \sum_{k=1}^{m} \log(X_{(k)}) - \log(X_{(m+1)}) \right]^{-1}. \qquad [6]$$

Empirically, all estimates $\hat{\alpha}$ fall in (1, 2), with one exception (Fig. 6). For cases, the lowest $\hat{\alpha} = 1.17$ was based on the 22 counties with the highest cases among the 2,214 US counties with reported cases on 1 April 2020. For cases, all remaining $\hat{\alpha}$ are based on 29 to 32 counties and fall between 1.26 (in May 2020) and 1.62 (in July 2020). For deaths, the lowest estimate of $\hat{\alpha} = 0.99$ on 1 April 2020 was based on only six counties, the top 1% of the 573 counties in the United States with reported deaths. All remaining estimates of $\hat{\alpha}$ for deaths are based on 15 to 31 counties and fall between 1.49 (in May 2020) and 1.93 (in January 2021). When the Pareto component of a lognormal–Pareto distribution has $\alpha \in (1, 2)$, the whole distribution has a finite mean and an infinite variance.

These results pose the challenge of reconciling a Pareto distribution of the largest counts, having tail index α in $(1, 2)$, with TL, having slope or exponent $b \approx 2$. The simulations and mathematical analyses in *SI Appendix* (Theorem 3) meet this challenge.

5. Summary of Simulations and Mathematical Analyses in *SI Appendix*

SI Appendix describes simulations that give evidence that TL with slope two holds for samples with the largest means drawn from heavy-tailed distributions. One simulation assumes independent observations, which is mathematically convenient but empirically implausible. A second simulation considers moderately and highly dependent observations. *SI Appendix* then states and proves three theorems inspired by the simulations and gives an example. These results are the mathematical heart of the paper. We then review our main theoretical findings and some of their limitations.

6. Discussion

Here we review and discuss our main empirical findings and some of their limitations. We also propose a practical consequence of our findings for planning the management of COVID-19

cases and deaths, especially in largely unvaccinated populations. *SI Appendix* gives some examples of other empirical data that could be, but have not yet been, analyzed using our approach in this paper.

We have demonstrated striking regularities in the variability among counties within states of reported counts of cumulative SARS-CoV-2 cases and cumulative COVID-19 deaths. In the United States, on the first day of each of 15 mo from April 2020 through June 2021, omitting the first few months of the pandemic, the variance of cases and, separately, deaths is nearly proportional to the square of the mean count of cases or deaths (across the counties within each state or other primary administrative unit of the United States). The approximately power-law relationship of the variance to the mean illustrates TL, a widespread pattern in ecology and epidemiology. To our knowledge, this pattern has not been recognized previously for SARS-CoV-2 cases and COVID-19 deaths.

The estimated slope b of TL closely approximates two after the first few months, for both cases and deaths. The slope $b = 2$ is the only positive value of b such that the coefficient of variation (SD divided by mean) is the same for all samples. Only for slope $b = 2$ is TL Eq. **2** scale invariant, with the same parameters a, b regardless of the scale on which observations are measured (formally, if $\sigma_X^2 = a\mu_X^2$ and $c > 0$, then $\sigma_{cX}^2 = a\mu_{cX}^2$). For example, if $b = 2$, then the parameters a, b are the same whether counts are measured in terms of individuals or millions of individuals.

We estimate that the largest 1% of counts of cases and deaths by county on each date are approximately Pareto distributed and that the upper tail of the empirical survival curve of counts by county has a tail index α between one and two. If that is correct, the underlying distributions of cases and deaths, except for, possibly, deaths in April 2020, have finite mean and infinite variance. This finding has implications for planning prevention and care: Facility and resource planning should prepare for rare but extremely high counts. No single county, state, region, or country can prepare in isolation for unboundedly high counts. Cooperative exchanges of support should be planned cooperatively.

Beare and Toda (20) analyzed *The New York Times* cumulative counts of cases by county on 31 March 2020, the day before the first observation we use here. The left portion of their survival curve, $\log \Pr\{X > x\}$ as a function of $\log x$, is concave, and hence not Pareto: Figure 1 of ref. 20 qualitatively resembles each panel in our Fig. 5. They found that the upper tail of the distribution of the number of cases for the top 6.2% of counties by number of cases reasonably approximated a straight line (on log–log coordinates) with estimated tail index 0.930 (SE 0.081). Hence their estimated tail index did not differ from one but was clearly less than two. Our estimate of the tail index of the top 1% of counties by number of cases on 1 April 2020 (Fig. 5, first panel), lies between 1.1 and 1.2, not far from the results of Beare and Toda (20). Our calculations were completed before we learned of ref. 20. In all months (except for deaths in April 2020), our estimated tail index of cases is greater than one but less than two.

Heavy-tailed distributions have also been used in modeling COVID-19 superspreader events, in which a primary infected individual infects an exceptionally large number of secondary individuals. That different use is not pursued here.

Our data analysis suffers from multiple limitations, some due to the data used and some due to our analyses. For example, we made no effort to estimate or correct underreporting of cases or deaths. We did not examine case fatality rates. We did not relate counts to the population at risk, whether treated as a population total or adjusted for age structure.

6.1. Data, Materials, and Software Availability. Previously published data were used for this work (https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv).

ACKNOWLEDGMENTS. J.E.C.'s research was partially supported by Columbia University's Earth Institute. R.A.D.'s research was partially supported by NSF Grant DMS 2015379 to Columbia University. G.S.'s research was partially supported by NSF Grant DMS-2015242 to Cornell University.

Author affiliations: [a]Laboratory of Populations, The Rockefeller University & Columbia University, New York, NY 10065; [b]Earth Institute, Columbia University, New York, NY 10027; [c]Department of Statistics, Columbia University, New York, NY 10027; [d]Department of Statistics, University of Chicago, Chicago, IL 60637; and [e]School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14853

1. Worldometer, *Countries where COVID-19 has spread*. https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/. Accessed 9 May 2022.
2. New York Times, *nytimes covid-19-data public*. https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv. Accessed 19 June 2021.
3. S. K. Bar-Lev, P. Enis, Reproducibility and natural exponential families with power variance functions. *Ann. Stat.* **14**, 1507–1522 (1986).
4. S. K. Bar-Lev, O. Stramer, Characterizations of natural exponential families with power variance functions by zero regression properties. *Probab. Theory Relat. Fields* **76**, 509–522 (1987).
5. M. Davidian, R. J. Carroll, Variance function estimation. *J. Am. Stat. Assoc.* **82**, 1079–1091 (1987).
6. Z. Eisler, I. Bartos, J. Kertész, Fluctuation scaling in complex systems: Taylor's law and beyond. *Adv. Phys.* **57**, 89–142 (2008).
7. R. A. Taylor, *Taylor's Power Law: Order and Pattern in Nature* (Academic, 2019).
8. J. E. Cohen, Stochastic population dynamics in a Markovian environment implies Taylor's power law of fluctuation scaling. *Theor. Popul. Biol.* **93**, 30–37 (2014).
9. J. Aitchison, J. Brown, *The Lognormal Distribution with Special Reference to Its Uses in Economics* (Cambridge University Press, Cambridge, 1957).
10. S. Goh, H. W. Kwon, M. Y. Choi, Discriminating between Weibull distributions and log-normal distributions emerging in branching processes. *J. Phys. A Math. Theor.* **47**, 225101 (2014).
11. M. K. Tippett, J. E. Cohen, Tornado outbreak variability follows Taylor's power law of fluctuation scaling and increases dramatically with severity. *Nat. Commun.* **7**, 10668 (2016).
12. W. W. Badger, "An entropy-utility model for the size distribution of income" in *Mathematical Models as a Tool for the Social Sciences*, B. J. West, Ed. (Gordon and Breach, New York, 1980), pp. 87–120.
13. E. W. Montroll, M. F. Shlesinger, Maximum entropy formalism, fractals, scaling phenomena, and 1 $/f$ noise: A tale of tails. *J. Stat. Phys.* **32**, 209–230 (1983).
14. W. Souma, "Physics of personal income" in *Empirical Science of Financial Fluctuations: The Advent of Econophyscs*, H. Takayasu, Ed. (Springer Japan KK, Tokyo, 2002), pp. 342–352.
15. K. Cooray, M. M. Ananda, Modeling actuarial data with a composite lognormal-Pareto model. *Scand. Actuar. J.* **2005**, 321–334 (2005).
16. D. P. M. Scollnik, On composite lognormal-Pareto models. *Scand. Actuar. J.* **2007**, 321–334 (2007).
17. M. Bee, Estimation of the lognormal-Pareto distribution using probability weighted moments and maximum likelihood. *Commun. Stat. Comput* **44**, 2040–2060 (2015).
18. B. M. Hill, A simple general approach to inference about the tail of a distribution. *Ann. Stat.* **3**, 1163–1174 (1975).
19. L. de Haan, A. Ferreira, *Extreme Value Theory: An Introduction* (Springer, New York, 2006).
20. B. K. Beare, A. A. Toda, On the emergence of a power law in the distribution of COVID-19 cases. *Physica D* **412**, 132649 (2020).
21. X. Gabaix, R. Ibragimov, Rank minus 1/2: A simple way to improve the OLS estimation of tail exponents. *J. Bus. Econ. Stat.* **29**, 24–39 (2011).