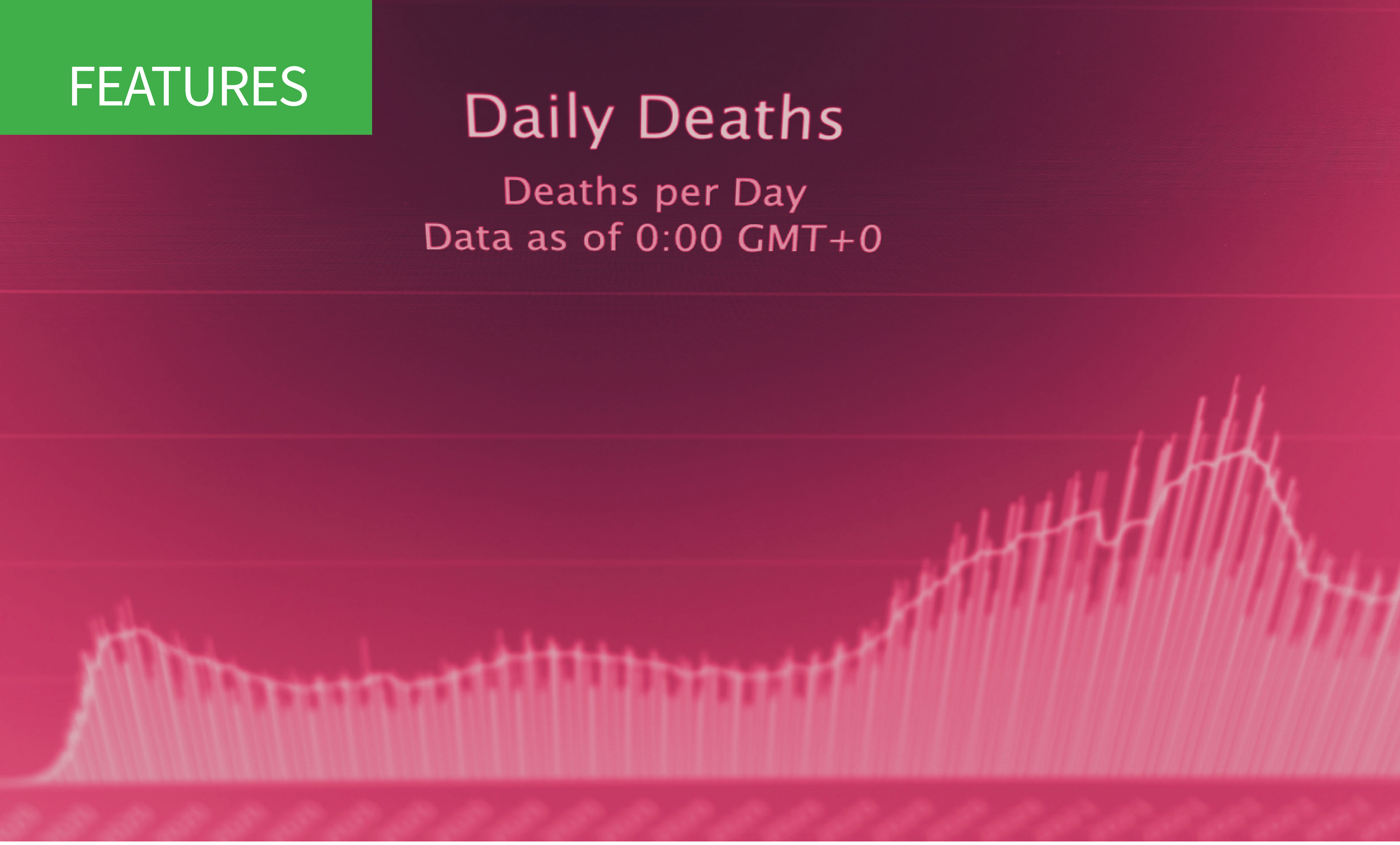

Underdispersion: A statistical anomaly in reported Covid data

Throughout the Covid-19 pandemic, we have become used to seeing daily numbers of cases and deaths go up and down. But in some countries, the reported numbers show very little movement over days and weeks – they are "underdispersed", says **Dmitry Kobak**, and this may be a sign that all is not right with the data

Since the beginning of the Covid-19 pandemic, countries around the world have been reporting key statistics each day, including the numbers of new Covid cases and deaths. These numbers are collated into dashboards, such as the one maintained by the World Health Organization (WHO; covid19.who.int), and they appear in countless media stories.

Yet it is well known that for many countries the reported numbers of cases and deaths can be gross underestimations, due to insufficient testing capacity or in some cases perhaps even purposeful misdiagnosing or misreporting of Covid infections.[1] In some countries, the reported numbers of Covid deaths are much lower than estimates of *excess mortality* – that is, deaths from all causes that exceed some baseline level of mortality – suggesting that the daily reported numbers are unreliable and incomplete.[2]

We have found that in some instances the reported numbers exhibit a statistical anomaly called *underdispersion*. Here, we explain what this anomaly is, how to test for it, and argue that it may be indicative of data tampering and deliberate obfuscation.

Dmitry Kobak is a research scientist at Tübingen University, Germany.

gnepphoto/Bigstock.com

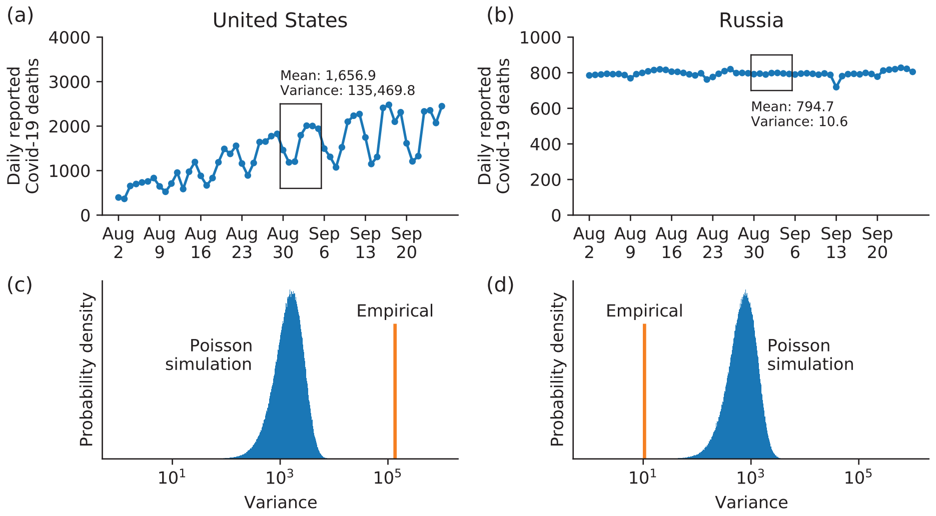

Figure 1: (a) Reported number of daily Covid-19 deaths in the United States over 8 weeks in August–September 2021. Ticks on the horizontal axis denote Mondays. The first week of September is highlighted. (b) The same for Russia. (c) The observed variance of the reported Covid-19 deaths in the first week of September in the USA (orange), and the distribution of variances expected if deaths followed the Poisson distribution with the observed mean (blue). (d) The same for Russia.

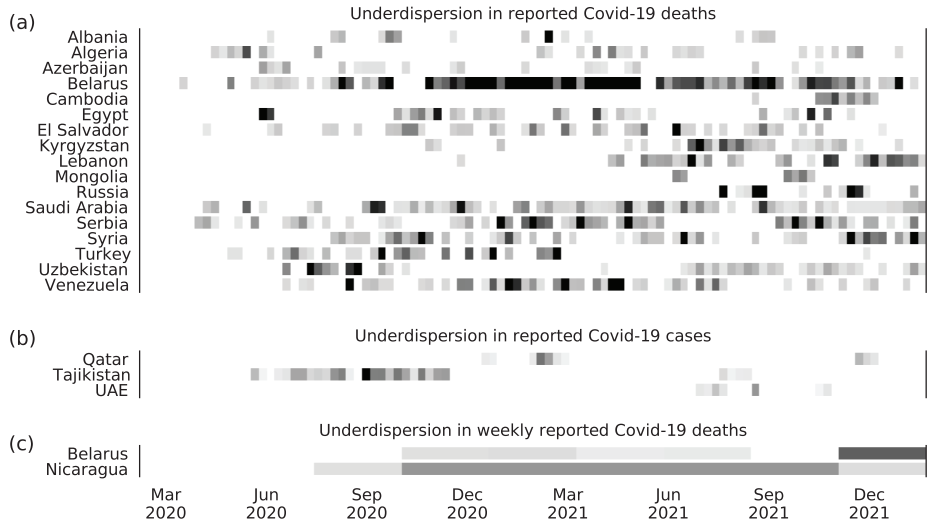

Figure 2: (a), (b) Screening all 237 countries and territories in the WHO data set for underdispersion in daily reported Covid-19 deaths and cases between 3 March 2020 and 30 January 2022 (100 weeks). Grey rectangles denote weeks (from Monday to Sunday) showing statistically significant ($p \leq 0.05$) underdispersion compared to the Poisson distribution with the same mean. Only countries with at least 15 significant weeks in total or at least four significant weeks in a row are shown. Shades of grey denote the mean divided by the variance in each week (black: 20 or higher). (c) Screening for underdispersion in weekly reported Covid-19 deaths in blocks of 10 weeks. Only countries with at least five significant blocks are shown.

What is underdispersion?

We will start with the number of Covid-19 deaths reported by the United States during August and September 2021 (Figure 1(a); here and below we use data distributed by the WHO). This time series shows several prominent features. First, the number of deaths is growing, corresponding to a rising wave of infections. Second, the daily numbers vary quite noticeably over the days of each week. Third, on top of the slow monotonic growth and day-of-week variations, there are additional random fluctuations. Similar patterns are observed in many other countries, which is why these data are typically presented in a smoothed form, such as weekly averages.

Some countries, however, do not show any such patterns. For example, the number of reported Covid deaths in Russia over the same time period stayed almost constant at just below 800 deaths per day (Figure 1(b)). Such lack of variation seems inconsistent with the random nature of the data-generating process: people get infected randomly, and the disease progresses independently in each patient.

Now, of course, there is no definitive answer to the question of how much variation we should expect to see in daily reported Covid deaths. Deaths will depend on a variety of factors, including the quality of health care, the availability of vaccines, and the age of those infected, while the reporting of deaths will be affected by the working patterns of staff in hospitals, public health bodies or government ministries, and by reporting procedures – for example, the USA reports Covid-19 deaths by date of registration, whereas some other countries report them by date of death. However, we can get a lower bound on the expected variation using a probability distribution known as the Poisson distribution.

If every citizen in a large country independently had the same small probability of dying from Covid-19 on any given day, daily deaths would follow the Poisson distribution. This remains true if probabilities differ between subpopulations (e.g., depending on geographical region or a person's sex or age), as long as each death is statistically independent. The Poisson distribution has the property that its mean equals its variance, meaning that if 800 people were to die per day on average, then the Poisson variance would also be 800.

In reality, the assumption that Covid deaths are independent does not hold: for example, one superspreading event can lead to an increase in infections or deaths on a particular day, and so the observed day-to-day variation will become *larger* than expected by the Poisson law. But the Poisson distribution serves as a useful benchmark: if the day-to-day variation is *smaller* than expected by the Poisson law, it strongly suggests that the time series was not truly random.

We can turn this argument into a formal statistical test. Taking the Russian data from ▶

the first week of September (daily deaths: 792, 795, 790, 798, 799, 796, 793; Figure 1(b), black box), we can compute the average number of daily deaths (794.7) and the observed variance (10.6; unbiased estimate). The observed variance is clearly much lower than the expected Poisson variance, but it is always possible – if unlikely – that the variance of Poisson random numbers could be as low as this. To find out just how unlikely, we can run a simulation. Simulating 1 million random numbers from the Poisson distribution with the same mean (794.7), we count the number of times the variance turns out smaller than the observed variance (10.6). This happened only 7 times (Figure 1(d)), corresponding to a one-sided *p*-value of 0.000007. The null hypothesis here is that the data are Poisson (or overdispersed), and the alternative hypothesis is that they are underdispersed. The test provides strong evidence that the Russian data were indeed underdispersed. While this conclusion may seem obvious in this particular case, having a formal statistical test will allow us to screen all countries for evidence of underdispersion.

For comparison, we apply the same procedure to the United States data from the same week (daily deaths: 1,461, 1,185, 1,202, 1,795, 2,010, 2,003, 1,942; Figure 1(a), black box). Here the average is 1,656.9 and the variance exceeds 130,000 (remember that variance is average *squared* deviation from the average). Generating 1 million random numbers from the Poisson distribution with the same mean, we never obtained as large a variance (Figure 1(c)), corresponding to the one-sided *p*-value equal to 1. This is not surprising: the real data are clearly overdispersed compared to Poisson, due to day-of-week fluctuations, epidemic growth, and other possible sources of variability. As previously noted, the USA reports Covid-19 deaths by date of registration, but we observed Poisson overdispersion in many other countries that report Covid-19 deaths by date of death, such as Belgium, Spain, and Sweden.

Screening for anomalies

With this statistical tool in hand, we were able to test for underdispersion the entire data set of Covid-19 cases and deaths, as collated by the WHO. This data set contains data on 237 countries and territories, and we analysed each week (from Monday to Sunday) between 3 March 2020 and 30 January 2022 (100 weeks in total). In each week, we deemed underdispersion to be statistically significant whenever we obtained $p \leq 0.05$ (using 1,000 Poisson simulations). We thus conducted 237 × 100 × 2 separate statistical tests. To mitigate the multiple testing problem, we considered a country as showing statistically significant underdispersion if it had either at least 15 weeks with statistically significant underdispersion, or at least four weeks in a row. The probability of that happening under the null hypothesis is 0.0007, so even when testing 237 countries, we should expect to see almost no false positives.

Seventeen countries came out with statistically significant underdispersion in reported Covid-19 deaths (Figure 2(a)), and three more countries showed statistically significant underdispersion in reported Covid-19 cases (Figure 2(b)). As we discuss below, for many of these countries there is strong independent evidence that Covid-19 deaths have been undercounted, suggesting that the underdispersion test yielded a sensible list of countries to focus on. Some of these countries have been previously highlighted by a related analysis of Covid cases.[3]

Seventeen countries showed underdispersion in reported deaths, and three more in reported cases

We additionally tested all 85 Russian federal regions as well as all 60 public health jurisdictions in the United States. In Russia, 82 regions out of 85 were flagged for underdispersion, with many regions showing statistically significant underdispersion during long periods of time, much longer than the 12 weeks that we obtained for Russia as a whole (Figure 2(a)). In contrast, not a single US jurisdiction showed evidence of underdispersion.

Furthermore, we tested reported Covid-19 numbers for underdispersion across weeks (Figure 2(c)). We summed reported Covid-19 cases and deaths within each week, and applied the same procedure as above to test for underdispersion of weekly counts over windows of 10 consecutive weeks (100/10 = 10 windows in total). Only two countries, Belarus and Nicaragua, showed statistically significant underdispersion in at least five windows (the probability of this happening

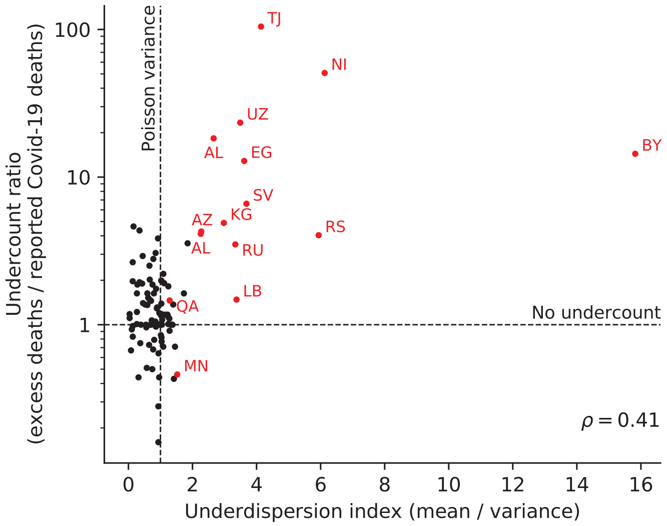

Figure 3: Relationship between the underdispersion index (ratio of observed mean to observed variance; averaged over 100 weeks or ten 10-week windows; maximum over four such average values, obtained for daily and weekly numbers of deaths and cases; we added 0.1 to the observed variances to avoid division by zero) and the undercount ratio (ratio of excess deaths to the reported deaths;[2] for countries with negative or near-zero excess mortality, here we set the undercount ratio to 1). Pearson correlation 0.41 (n = 113; all countries present in both the World Mortality Dataset and the WHO data set). Red dots denote countries listed in Figure 2, labelled with their ISO codes. Note that many countries do not have any estimate of the undercount ratio due to lack of excess mortality data; those countries are not shown.

under the null hypothesis was below 0.0001). Nicaragua is a particularly telling case as it has been reporting exactly one death per week since the beginning of 2021. This clearly implausible lack of variation is correctly picked up by our test.

Underdispersion and excess mortality

As mentioned earlier, in many countries estimates of excess deaths during the pandemic are much higher than daily reported Covid deaths. How much higher? Well, if we divide excess deaths by Covid deaths, we derive the *undercount ratio*.[2] For countries with reliable reporting, this ratio is around 1 or often even below 1. But for many countries listed in Figure 2, the undercount ratio is much larger than that: at the time of writing (31 January 2022), it is estimated to be above 10 in Algeria, Belarus, Egypt, Nicaragua, Uzbekistan, and Tajikistan, and between 3.5 and 10 in Albania, Azerbaijan, El Salvador, Kyrgyzstan, Russia, and Serbia (github.com/dkobak/excess-mortality). For Turkey, the undercount ratio has been estimated at approximately 3, based on incomplete excess mortality data (gucluyaman.com/excess-mortality-in-turkey). For Syria, the undercount ratio in its capital Damascus has been estimated at approximately 17, based on obituary notifications.[4]

Large undercount ratios could be due to many factors, including limited Covid testing or reporting capacity (such as due to ongoing conflict, which may be the case in Syria). They may also be due to misreporting or misdiagnosing of Covid deaths, whether done innocently or deliberately.[2] But a large undercount ratio together with a high level of underdispersion may suggest deliberate misreporting.

Only three countries flagged for underdispersion in Figure 2 have moderate undercount ratios, namely Lebanon (1.5), Mongolia (less than 1), and Qatar (1.5). For Cambodia, Saudi Arabia, the United Arab Emirates (UAE), and Venezuela, no data on excess deaths are available so far. Our results suggest that the undercounts there may be large.

Overall, of the 10 countries with the highest undercount ratios in the World Mortality Dataset at the time of writing,[2] the top eight demonstrated statistically significant underdispersion in our analysis. To quantify the relationship between undercount and underdispersion, we defined the *underdispersion index* as the ratio of the observed mean to the observed variance (averaged over 100 weeks or ten 10-week windows; for each country we took the maximum over four such average values, obtained for daily and weekly numbers of deaths and cases). For Poisson data, the underdispersion index should be around 1. The correlation of underdispersion index to undercount ratio was 0.41 (Figure 3). Most importantly, large values of the underdispersion index were always associated with high undercount. This suggests a large undercount in several countries with large values of the underdispersion index but no currently available estimates of excess mortality (e.g., Saudi Arabia, Syria, Turkey, Venezuela).

Poisson underdispersion provides a simple and useful test to detect one kind of reporting anomaly

A red flag

Underdispersion in reported Covid-19 cases or deaths cannot serve as definitive proof of data tampering, as it can in principle arise through other mechanisms. For example, a country may reach the limit of its testing capacity and so report the same maximally possible number of new cases every day. However, this particular hypothetical reason is unlikely to explain the observed patterns: most of the underdispersion was observed in reported deaths while the number of reported cases was much larger (so Covid tests could not have been the limiting factor). Also, in many cases, underdispersion was observed during weeks where the reported numbers were smaller than the same country had reported in the past, again suggesting that testing or reporting capacity could not have been reached.

We believe that the most likely explanation for the observed underdispersion patterns is deliberate data tampering, whereby a country or region does not report publicly the same values as were internally obtained. It is important to stress that such tampering may not necessarily have malicious intent (e.g., if somebody were "redistributing" the values across days). But, together with the evidence of underreporting coming from excess mortality, underdispersion is strongly suggestive of attempts to obfuscate.

It should also be stressed that if a country does not show evidence of underdispersion, this does not mean that Covid reporting there has been accurate. There may exist other statistical anomalies that could be indicative of misreporting. Furthermore, there is strong evidence that in many developing countries without reliable mortality tracking, the Covid undercount may be very high,[5] even though the reporting itself may be done honestly and without any anomalies. Despite these obvious limitations, here we argue that Poisson underdispersion provides a simple and useful test to detect one kind of reporting anomaly, and to highlight potentially unreliable data. ■

Code availability

Analysis code, together with the frozen data, is available at github.com/dkobak/covid-underdispersion.

Acknowledgements

The author thanks Stephan Glöckner, Ariel Karlinsky, Maxim Pshenichnikov, Boudewijn Roukema, and Oliver Watson for comments and suggestions, and Sergey Shpilkin for scraping and sharing the regional Russian data.

Disclosure statement

The author declares no conflicts of interest.

References

1. Kobak, D. (2021) Excess mortality reveals Covid's true toll in Russia. *Significance*, **18**(1), 16–19.
2. Karlinsky, A. and Kobak, D. (2021) Tracking excess mortality across countries during the COVID-19 pandemic with the World Mortality Dataset. *eLife*, **10**, e69336.
3. Roukema, B. F. (2021) Anti-clustering in the national SARS-CoV-2 daily infection counts. *PeerJ*, **9**, e11856.
4. Watson, O. J., Alhaffar, M., Mehchy, Z. *et al.* (2021) Leveraging community mortality indicators to infer COVID-19 mortality and transmission dynamics in Damascus, Syria. *Nature Communications*, **12**, 2394.
5. Whittaker, C., Walker, P. G. T., Alhaffar, M. *et al.* (2021) Under-reporting of deaths limits our understanding of true burden of covid-19. *British Medical Journal*, **375**, n2239.