PERSPECTIVE
https://doi.org/10.1038/s41567-020-0921-x

nature physics



# Tail risk of contagious diseases

Pasquale Cirillo[1]✉ and Nassim Nicholas Taleb[2]✉

**The COVID-19 pandemic has been a sobering reminder of the extensive damage brought about by epidemics, phenomena that play a vivid role in our collective memory, and that have long been identified as significant sources of risk for humanity. The use of increasingly sophisticated mathematical and computational models for the spreading and the implications of epidemics should, in principle, provide policy- and decision-makers with a greater situational awareness regarding their potential risk. Yet most of those models ignore the tail risk of contagious diseases, use point forecasts, and the reliability of their parameters is rarely questioned and incorporated in the projections. We argue that a natural and empirically correct framework for assessing (and managing) the real risk of pandemics is provided by extreme value theory (EVT), an approach that has historically been developed to treat phenomena in which extremes (maxima or minima) and not averages play the role of the protagonist, being the fundamental source of risk. By analysing data for pandemic outbreaks spanning over the past 2500 years, we show that the related distribution of fatalities is strongly fat-tailed, suggesting a tail risk that is unfortunately largely ignored in common epidemiological models. We use a dual distribution method, combined with EVT, to extract information from the data that is not immediately available to inspection. To check the robustness of our conclusions, we stress our data to account for the imprecision in historical reporting. We argue that our findings have significant implications, including on the extent to which compartmental epidemiological models and similar approaches can be relied upon for making policy decisions.**

We examine the distribution of fatalities from major pandemics in history (spanning about 2,500 years), and build a statistical picture of their tail properties. Using tools from extreme value theory (EVT), we show that the distribution of the victims of infectious diseases is extremely fat-tailed, more than what one could be led to believe from the outset. Our goal is not, at this stage, to explain the emergence of these tail properties, but several plausible explanations are already available[1–3].

A non-negative continuous random variable $X$ is 'fat-tailed' if its survival function $S(x) = P(X \geq x)$ decays as a power law $x^{-1/\xi}$ the more we move into the tail, that is, for $x$ growing towards the right endpoint of $X$. More technically, $X$ has a regularly varying survival function, that is $S(x) = L(x)\, x^{-1/\xi}$, where $L(x)$ is a slowly varying function, such that $\lim_{x\to\infty} \frac{L(cx)}{L(x)} = 1$ for $c > 0$ (refs. [2,3]). The parameter $\xi > 0$ is known as the tail parameter, and it governs the 'fatness' of the tail: the larger $\xi$, the fatter the tail. Moreover, $\xi$ determines the existence of moments, which include quantities such as the mean, the variance or the skewness. In fact, the moment of order $p$ exists, that is $E[X^p] < \infty$, if and only if $\xi < 1/p$. For example, when $\xi \geq 1/2$, the second moment $E[X^2]$ is not finite, and so the variance, which is nothing more than the second central moment, does not exist; just recall that $\mathrm{Var}(X) = E[X^2] - E[X]^2$. In some literature, for example, ref. [4], the tail index is re-parameterised as $\alpha = 1/\xi$, and its interpretation is naturally reversed.

In EVT, fat-tailed distributions are said to be in the maximum domain of attraction of the Fréchet distribution[2,3], representing the potentially most-dangerous generators of extreme risks. In this class, we naturally find all the different versions of the Pareto distribution, as well as the Cauchy, the log-gamma and all stable distributions with stability parameter $\alpha < 2$, to name a few. Well-known distributions such as the Gaussian and the exponential do not belong to this group, and are therefore considered thin-tailed. There are then distributions such as the log-normal, which are extremely tricky. Despite being theoretically in the same EVT class as the Gaussian distribution, the so-called Gumbel class, the log-normal distribution can be compared to a wolf in sheep's clothing, showing an erratic behaviour depending on its parameters[5,6].

While it is known that fat tails represent a common—yet often ignored[5]—regularity in many fields of science and knowledge[4], to the best of our knowledge, only war casualties and operational risk losses show a behaviour[7–9] as erratic and wild as the one we observe for pandemic fatalities.

The core of the problem is shown in Fig. 1, with the maximum-to-sum plot[9] of the number of pandemic fatalities in history, using the data collected in Table 1. Such a plot relies on a simple consequence of the law of large numbers: for a sequence $X_1, X_2, \ldots, X_n$ of non-negative independent and identically distributed random variables, if $E[X^p] < \infty$ for $p = 1, 2, 3\ldots$, then $R_n^p = M_n^p / S_n^p \to 0$ almost surely for $n \to \infty$, where $S_n^p = \sum_{i=1}^{n} X_i^p$ is the partial sum of order $p$, and $M_n^p = \max(X_1^p, \ldots, X_n^p)$ the corresponding partial maximum. Figure 1 clearly shows that no finite moment is likely to exist for the number of victims in pandemics, as the $R_n^p$ ratio does not converge to 0 for $p = 1, 2, 3, 4$, no matter how many data points we use. Such a behaviour suggests that the victims' distribution has such a fat right tail that not even the first theoretical moment is finite. We are therefore dealing with a phenomenon for which observed quantities such as the sample average and standard deviation are meaningless for inference.

However, this does not mean that pandemic risk is infinite and there is nothing we can do or model to get a handle on the problem. Using the methodology we have developed to study war casualties[7,10], it is possible to extract useful information from the data, and quantify the large (yet finite) risk of pandemic diseases. In fact, our method provides rough estimates for quantities not immediately observable from the data.

## The tail-wags-the-dog effect

The central point we wish to convey is the following: the more fat-tailed a statistical distribution, the more the 'tail wags the dog'. That is to say, more statistical information resides in the extremes

[1]Applied Probability Group, Delft University of Technology, Delft, The Netherlands. [2]Tandon School of Engineering, New York University, New York, NY, USA. ✉e-mail: p.cirillo@tudelft.nl; nnt1@nyu.edu

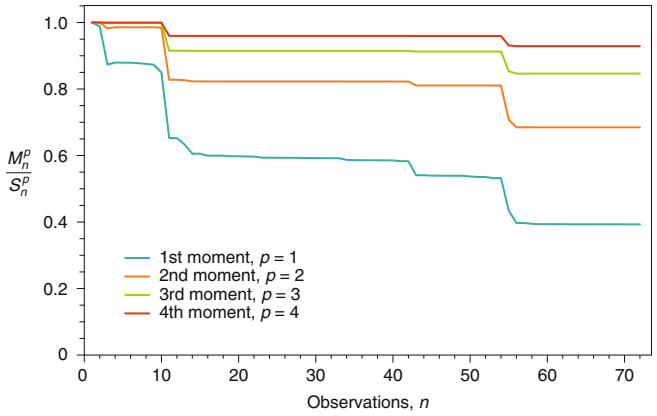


**Fig. 1 | Maximum to sum plot.** Maximum to sum plot (MS plot) of the average death numbers in pandemic events in history, based on the data presented in Table 1. The quantities $M_n^p$ and $S_n^p$ are the partial maximum and the partial sum of order $p$, respectively. The plot suggests that all moments, including the mean, could be infinite.

and less in the 'bulk'—the events of high frequency—where it becomes almost noise. Under fat tails, the law of large numbers works slowly, and moments—even when they exist—may become uninformative and unreliable[5]. All this makes EVT the most effective and robust approach for risk management purposes, even with relatively small datasets like ours (see Table 1).

The presence of a fat right tail in the distribution of pandemic fatalities has the following policy implications, useful in the wake of the COVID-19 pandemic.

First, it should be evident that it is not appropriate to compare fatalities from multiplicative infectious diseases (fat-tailed, like a Pareto distribution) to those from car accidents, heart attacks or falls from ladders (thin-tailed, like a Gaussian). This remains a common (and costly) error in policy making, and in both the decision sciences and the journalistic literature. Some research papers even criticise the wider public's 'paranoia' with respect to pandemics, not appreciating that such a paranoia is merely responsible (and realistic) risk management in front of potentially destructive events. The main problem is that those articles—often relied upon for policy making—consistently use the wrong thin-tailed distributions, underestimating tail risk, so that every conservative or preventative reaction is bound to be considered an overreaction.

Second, epidemiological models such as the susceptible–infectious–recovered (SIR) differential equations[11], sometimes supplemented with simulation experiments[12], while useful for scientific discussions for the bulk of the distributions of infections and deaths, or for understanding the dynamics of events after they have happened, should not be used for precautionary risk management, which should focus on maxima and tail exposures instead. It is not rigorous to use naive (but reassuring) statistics, such as the expected average outcome of compartmental models, or one or more point estimates, as a motivation for policies. Owing to the compounding effect of parameter uncertainty, the tail-wags-the-dog effect easily invalidates both point estimates and scenario analyses. However, it is encouraging to note that the impact of parameter uncertainty on the scenarios generated by epidemiological models has recently started to be examined more carefully[13].

Extreme value theory is a natural candidate to handle pandemics. It was developed as a means to cope with maxima[14], and it has subsequently evolved to deal with tail risk in a robust way, even with a limited number of observations and their associated uncertainty[3]. In the Netherlands, for example, EVT has been used to get a handle on the distribution of the maxima—and not the average—of sea levels in order to build dams and dykes high and strong enough for the safety of its citizens[2].

Finally, EVT-based risk management is compatible with the (non-naive) precautionary principle[15], which should be the leading driver for policy decisions under jointly systemic and extreme risks.

## Data and descriptive statistics

We investigate the distribution of deaths from the major epidemic and pandemic diseases of history, from 429 BC until the present. The data are available in Table 1, together with their sources, and only refer to events with more than 1,000 estimated victims, for a total of 72 observations. As a consequence, potentially high-risk diseases such as the Middle East Respiratory Syndrome (MERS) do not appear in our collection. All diseases whose end year is 2020 are to be considered ongoing, as is the case for the current COVID-19 pandemic.

We use three estimates of the reported cumulative death toll: minimum, average and maximum. When the three numbers coincide in Table 1, our sources simply do not provide intervals for the estimates. Since we are well aware of the volatility and possible unreliability of historical data[10,16], in Section 5 we deal with the issue by perturbing and omitting observations.

In order to compare fatalities with respect to the coeval population (that is, the relative impact of pandemics), the 'Rescaled' column in Table 1 provides the rescaled version of the 'Average estimate' column, using the information in the 'Population' column[17–19]. For example, the Antonine plague of 165–180 killed an average of 7.5 million people, that is to say 3.7% of the coeval world population of 202 million people. Using today's population, such a number would correspond to about 283 million deaths, a terrible hecatomb, killing more people than World War II.

We restrict our attention to the actual average estimates in Table 1, but all our findings and conclusions also hold for the lower, the upper and the rescaled estimates as well.

Figure 2a shows the histogram of the average numbers of deaths in the 72 large contagious events. The distribution appears highly skewed and possibly fat-tailed. The numbers are as follows: the sample average is 4.9 million, while the median is 76 thousand, compatible with the skewness observable in Fig. 2. The 90% quantile is 6.5 million and the 99% quantile is 137.5 million. The sample standard deviation is 19 million.

Using common graphical tools for fat tails[3,6], in Fig. 2b we show the log–log plot (also known as the Zipf plot) of the empirical survival functions for the average victims over the diverse contagious events. In such a plot, possible fat tails can be identified in the presence of a linearly decreasing behaviour of the plotted curve. To improve interpretability, a naive linear fit is also proposed. Figure 2b also suggests the presence of fat tails.

The Zipf plot shows a necessary but not sufficient condition for fat tails[6]. Therefore, in Fig. 2c we complement the analysis with a mean excess function plot (meplot). If a random variable $X$ is possibly fat tailed, its mean excess function $e_X(u) = E[X - u|X \geq u]$ should grow linearly in the threshold $u$, at least above a certain value identifying the actual power law tail[3]. In a meplot, where the empirical $e_X(u)$ is plotted against the different values of $u$, one thus looks for some (more or less) increasing trend, such as the one we observe in Fig. 2c.

A useful tool for the analysis of tails—when one suspects them of being fat—is the non-parametric Hill estimator[2,3]. For a collection $X_1, \ldots, X_n$, let $X_{n,n} \leq \ldots \leq X_{1,n}$ be the corresponding order statistics. We can then estimate the tail parameter $\xi$ as

$$\hat{\xi} = \frac{1}{k}\sum_{i=1}^{k} \log(X_{i,n}) - \log(X_{k,n}), 2 \leq k \leq n.$$

**Table 1 | The data set used for the analysis**

| Name | Start Year | End Year | Lower Est ($\times 10^3$) | Avg Est ($\times 10^3$) | Upper Est ($\times 10^3$) | Rescaled Avg Est ($\times 10^3$) | Population ($\times 10^6$) | Source |
|---|---|---|---|---|---|---|---|---|
| Plague of Athens | -429 | -426 | 75 | 88 | 100 | 13,376 | 50 | Ref. 22 |
| Antonine Plague | 165 | 180 | 5,000 | 7,500 | 10,000 | 283,355 | 202 | Ref. 22 |
| Plague of Cyprian | 250 | 266 | 1,000 | 1,000 | 1,000 | 37,227 | 205 | Ref. 22 |
| Plague of Justinian | 541 | 542 | 25,000 | 62,500 | 100,000 | 2,246,550 | 213 | Ref. 22 |
| Plague of Amida | 562 | 562 | 30 | 30 | 30 | 1,078 | 213 | Ref. 23 |
| Roman Plague of 590 | 590 | 590 | 10 | 20 | 30 | 719 | 213 | Ref. 22 |
| Plague of Sheroe | 627 | 628 | 100 | 100 | 100 | 3,594 | 213 | Ref. 24 |
| Plague of the British Isles | 664 | 689 | 150 | 175 | 200 | 6,290 | 213 | Ref. 22 |
| Plague of Basra | 688 | 689 | 200 | 200 | 200 | 7,189 | 213 | Ref. 23 |
| Japanese smallpox epidemic | 735 | 737 | 2,000 | 2,000 | 2,000 | 67,690 | 226 | Ref. 22 |
| Black Death | 1331 | 1353 | 75,000 | 137,500 | 200,000 | 2,678,283 | 392 | Ref. 22 |
| Sweating sickness | 1485 | 1551 | 10 | 10 | 10 | 166 | 461 | Ref. 22 |
| Smallpox Epidemic in Mexico | 1520 | 1520 | 5,000 | 6,500 | 8,000 | 107,684 | 461 | Ref. 22 |
| Cocoliztli Epidemic of 1545–1548 | 1545 | 1548 | 5,000 | 10,000 | 15,000 | 165,668 | 461 | Ref. 22 |
| 1563 London plague | 1562 | 1564 | 20 | 20 | 20 | 277 | 554 | Ref. 22 |
| Cocoliztli epidemic of 1576 | 1576 | 1580 | 2,000 | 2,250 | 2,500 | 31,045 | 554 | Ref. 22 |
| 1592–93 London plague | 1592 | 1593 | 20 | 20 | 20 | 275 | 554 | Ref. 22 |
| Malta plague epidemic | 1592 | 1593 | 3 | 3 | 3 | 41 | 554 | Ref. 22 |
| Plague in Spain | 1596 | 1602 | 600 | 650 | 700 | 8,969 | 554 | Ref. 22 |
| New England epidemic | 1616 | 1620 | 7 | 7 | 7 | 97 | 554 | Ref. 22 |
| Italian plague of 1629–1631 | 1629 | 1631 | 280 | 280 | 280 | 3,863 | 554 | Ref. 22 |
| Great Plague of Sevilla | 1647 | 1652 | 150 | 150 | 150 | 2,070 | 554 | Ref. 22 |
| Plague in Kingdom of Naples | 1656 | 1658 | 1,250 | 1,250 | 1,250 | 15,840 | 603 | Ref. 25 |
| Plague in the Netherlands | 1663 | 1664 | 24 | 24 | 24 | 306 | 603 | Ref. 22 |
| Great Plague of London | 1665 | 1666 | 100 | 100 | 100 | 1,267 | 603 | Ref. 22 |
| Plague in France | 1668 | 1668 | 40 | 40 | 40 | 507 | 603 | Ref. 22 |
| Malta plague epidemic | 1675 | 1676 | 11 | 11 | 11 | 143 | 603 | Ref. 22 |
| Great Plague of Vienna | 1679 | 1679 | 76 | 76 | 76 | 963 | 603 | Ref. 22 |
| Great Northern War plague outbreak | 1700 | 1721 | 176 | 192 | 208 | 2,427 | 603 | Ref. 26 |
| Great Smallpox Epidemic in Iceland | 1707 | 1709 | 18 | 18 | 18 | 228 | 603 | Ref. 22 |
| Great Plague of Marseille | 1720 | 1722 | 100 | 100 | 100 | 1,267 | 603 | Ref. 22 |
| Great Plague of 1738 | 1738 | 1738 | 50 | 50 | 50 | 470 | 814 | Ref. 22 |
| Russian plague of 1770–1772 | 1770 | 1772 | 50 | 50 | 50 | 470 | 814 | Ref. 22 |
| Persian Plague | 1772 | 1772 | 2,000 | 2,000 | 2,000 | 15,444 | 990 | Ref. 22 |
| Ottoman Plague Epidemic | 1812 | 1819 | 300 | 300 | 300 | 2,317 | 990 | Ref. 24 |
| Caragea's plague | 1813 | 1813 | 60 | 60 | 60 | 463 | 990 | Ref. 24 |
| Malta plague epidemic | 1813 | 1814 | 5 | 5 | 5 | 35 | 990 | Ref. 22 |
| First cholera pandemic | 1816 | 1826 | 100 | 100 | 100 | 772 | 990 | Ref. 22 |
| Second cholera pandemic | 1829 | 1851 | 100 | 100 | 100 | 772 | 990 | Ref. 22 |
| Typhus epidemic in Canada | 1847 | 1848 | 20 | 20 | 20 | 154 | 990 | Ref. 22 |
| Third cholera pandemic | 1852 | 1860 | 1,000 | 1,000 | 1,000 | 6,053 | 1,263 | Ref. 22 |
| Cholera epidemic of Copenhagen | 1853 | 1853 | 5 | 5 | 5 | 29 | 1,263 | Ref. 24 |
| Third plague pandemic | 1855 | 1960 | 15,000 | 18,500 | 22,000 | 111,986 | 1,263 | Refs. 22–24 |
| Smallpox in British Columbia | 1862 | 1863 | 3 | 3 | 3 | 18 | 1,263 | Ref. 24 |

Continued

**Table 1 | The data set used for the analysis (continued)**

| Name | Start Year | End Year | Lower Est (×10³) | Avg Est (×10³) | Upper Est (×10³) | Rescaled Avg Est (×10³) | Population (×10⁶) | Source |
|---|---|---|---|---|---|---|---|---|
| Fourth cholera pandemic | 1863 | 1875 | 600 | 600 | 600 | 3,632 | 1,263 | Ref. 24 |
| Fiji Measles outbreak | 1875 | 1875 | 40 | 40 | 40 | 242 | 1,263 | Ref. 24 |
| Yellow Fever | 1880 | 1900 | 100 | 125 | 150 | 757 | 1,263 | Ref. 27 |
| Fifth cholera pandemic | 1881 | 1896 | 9 | 9 | 9 | 42 | 1,654 | Ref. 22 |
| Smallpox in Montreal | 1885 | 1885 | 3 | 3 | 3 | 14 | 1,654 | Ref. 22 |
| Russian flu | 1889 | 1890 | 1,000 | 1,000 | 1,000 | 4,620 | 1,654 | Ref. 22 |
| Sixth cholera pandemic | 1899 | 1923 | 800 | 800 | 800 | 3,696 | 1,654 | Ref. 24 |
| China plague | 1910 | 1912 | 40 | 40 | 40 | 185 | 1,654 | Ref. 22 |
| Encephalitis lethargica pandemic | 1915 | 1926 | 1,500 | 1,500 | 1,500 | 6,930 | 1,654 | Ref. 22 |
| American polio epidemic | 1916 | 1916 | 6 | 7 | 7 | 30 | 1,654 | Ref. 22 |
| Spanish flu | 1918 | 1920 | 17,000 | 58,500 | 100,000 | 193,789 | 2,307 | Ref. 22 |
| HIV/AIDS pandemic | 1920 | 2020 | 25,000 | 30,000 | 35,000 | 61,768 | 3,712 | Ref. 24·27 |
| Poliomyelitis in USA | 1946 | 1946 | 2 | 2 | 2 | 5 | 2,948 | Ref. 22 |
| Asian flu | 1957 | 1958 | 2,000 | 2,000 | 2,000 | 5,186 | 2,948 | Ref. 22 |
| Hong Kong flu | 1968 | 1969 | 1,000 | 1,000 | 1,000 | 2,102 | 3,637 | Ref. 22 |
| London flu | 1972 | 1973 | 1 | 1 | 1 | 2 | 3,866 | Ref. 22 |
| Smallpox epidemic of India | 1974 | 1974 | 15 | 15 | 15 | 29 | 4,016 | Ref. 22 |
| Zimbabwean cholera outbreak | 2008 | 2009 | 4 | 4 | 4 | 5 | 6,788 | Ref. 22 |
| Swine Flu | 2009 | 2009 | 152 | 364 | 575 | 409 | 6,788 | Ref. 22 |
| Haiti cholera outbreak | 2010 | 2020 | 10 | 10 | 10 | 11 | 7,253 | Ref. 22 |
| Measles in D.R. Congo | 2011 | 1018 | 5 | 5 | 5 | 5 | 7,253 | Ref. 22 |
| Ebola in West Africa | 2013 | 2016 | 11 | 11 | 11 | 12 | 7,176 | Ref. 22 |
| Indian swine flu outbreak | 2015 | 2015 | 2 | 2 | 2 | 2 | 7,253 | Ref. 22 |
| Yemen cholera outbreak | 2016 | 2020 | 4 | 4 | 4 | 4 | 7,643 | Ref. 22 |
| 2018–19 Kivu Ebola epidemic | 2018 | 2020 | 2 | 2 | 3 | 2 | 7,643 | Ref. 22 |
| 2019-20 COVID-19 Pandemic | 2019 | 2020 | 117 | 133.5 | 150 | 50 | 7,643 | Ref. 20 |
| Measles in D.R. Congo | 2019 | 2020 | 5 | 5 | 5 | 5 | 7,643 | Ref. 24 |
| Dengue fever | 2019 | 2020 | 2 | 2 | 2 | 2 | 7,643 | Ref. 22 |

All estimates in thousands, apart from coeval population, which is expressed in millions. For COVID-19 (ref. [20], last update considered on 13 April 2020), the upper estimate includes the supposed number of Chinese victims (42,000) for some Western media.

In Fig. 2d, the estimate $\hat{\xi}$ is plotted against different values of $k$, creating the so-called Hill plot[3]. The plot suggests $\xi > 1$, in line with Fig. 1, further supporting the evidence of infinite moments.

Other graphical tools could be used and they would all confirm our basic point: we are in the presence of a fat right tail in the distribution of the victims of pandemic diseases. Furthermore, it is possibly a distribution with no finite moment (not even the mean).

## The dual distribution

As we observed for war casualties[7], the lack of moments for the distribution of pandemic victims is questionable. Since the distribution of victims is naturally bounded by the coeval world population, no disease can kill more people than those living on the planet at any given time. We are indeed looking at an apparently infinite-mean phenomenon, like in the case of war casualties[7,10] and operational risk[8].

Let $[L, H]$ be the support of the distribution of pandemic victims today (that is, the range of variation for the number of fatalities), with $L >> 0$ to ignore small events not officially definable as a pandemic[20]. For what concerns $H$, its value cannot be larger than the world population, and we can safely take today's estimates as the historical upper bound, that is, 7.7 billion people in 2020[19]. Evidently, $H$ is so large that the probability of observing values in its vicinity is in practice zero, and one always finds observations below a given $M << H < \infty$ (something like 150 million deaths using actual data). Thus, one could be fooled by data into ignoring $H$ and taking it as infinite, up to the point of believing in an infinite mean phenomenon, as Fig. 1 suggests. However, notice that a finite upper bound $H$—no matter how large it is—is not compatible with infinite moments[3], hence Fig. 1 risks being misleading.

In Fig. 3, the real tail of the random variable $Y$ with remote upper bound $H$ is represented by the dashed line. If one only observes values up to $M << H$, and more or less consciously ignores the existence of $H$, one could be fooled by the data into believing that the tail is actually the continuous one, the so-called apparent tail[8]. The tails are indeed indistinguishable for most cases, virtually in all

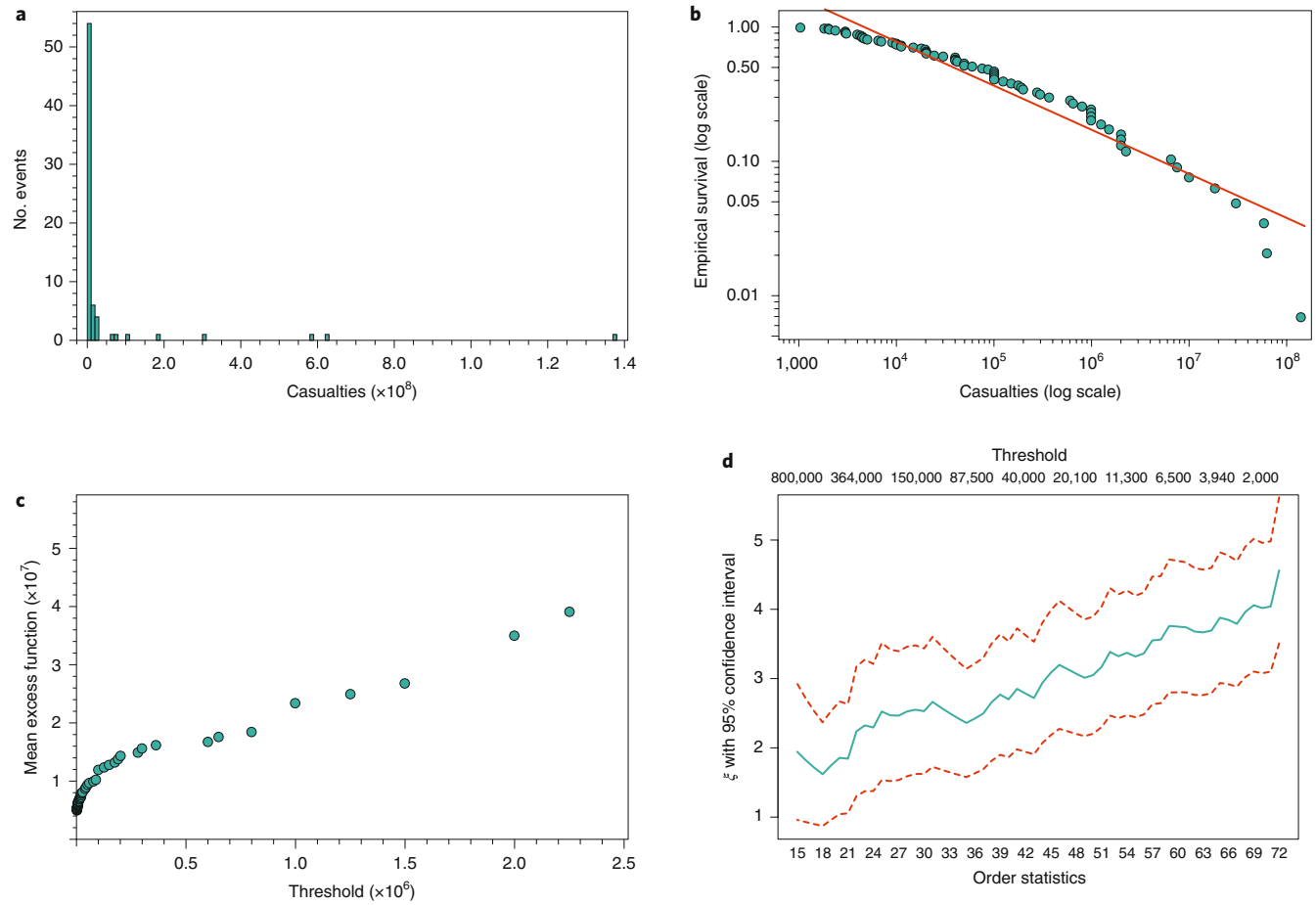


**Fig. 2 | Graphical analyses of the average number of deaths. a**, Histogram of the average number of deaths in the 72 contagious diseases of Table 1. **b**, Log–log plot of the empirical survival function (Zipf plot) of the actual average death numbers in Table 1. The red line represents a naive linear fit of the decaying tail. **c**, Mean excess function plot (meplot) of the average death numbers in Table 1. The plot excludes three points on the top right corner, consistently with the suggestions in ref. [9] about the exclusion of the more volatile observations. **d**, Hill plot of the average death numbers in Table 1, with 95% confidence intervals. Clearly $\xi > 1$, suggesting the non-existence of moments.

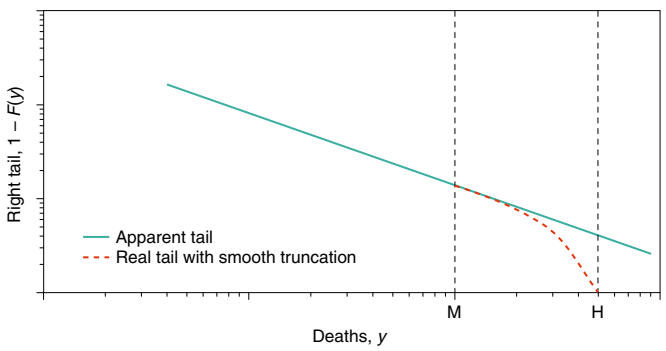


**Fig. 3 | The apparent and real tail.** Graphical representation (log–log plot) of what may happen if one ignores the existence of the finite upper bound *H*, since only *M* is observed.

finite samples, as the divergence is only clear in the vicinity of *H*. A bounded tail with very large upper limit is therefore mistakenly confused with an unbounded one, and no model will be able to tell the difference, even if epistemologically we are in two extremely different situations. This is the typical case in which critical reasoning, and the *a priori* analysis of the characteristics of the phenomenon under scrutiny, should precede any instinctive and uncritical fitting of the data.

A solution to this quandary is provided by the approach of refs. [7,8], which introduces the concept of dual data via a special log transformation. The basic idea is to find a way of matching naive extrapolations (apparently infinite moments) with appropriate modelling.

Let $L$ and $H$ be, respectively, the finite lower and upper bounds of a random variable $Y$, and define the function

$$\varphi(Y) = L - H\log\left(\frac{H-Y}{H-L}\right), \qquad (1)$$

from which it follows that

$$\varphi \in C^{\infty},$$

$$\varphi^{-1}(\infty) = H,$$

$$\varphi^{-1}(L) = \varphi(L) = L.$$

We then define $Z = \varphi(Y)$ as a new random variable with lower bound $L$ and an infinite upper bound. The transformation induced by $\varphi$ does not depend on any of the parameters of the distribution

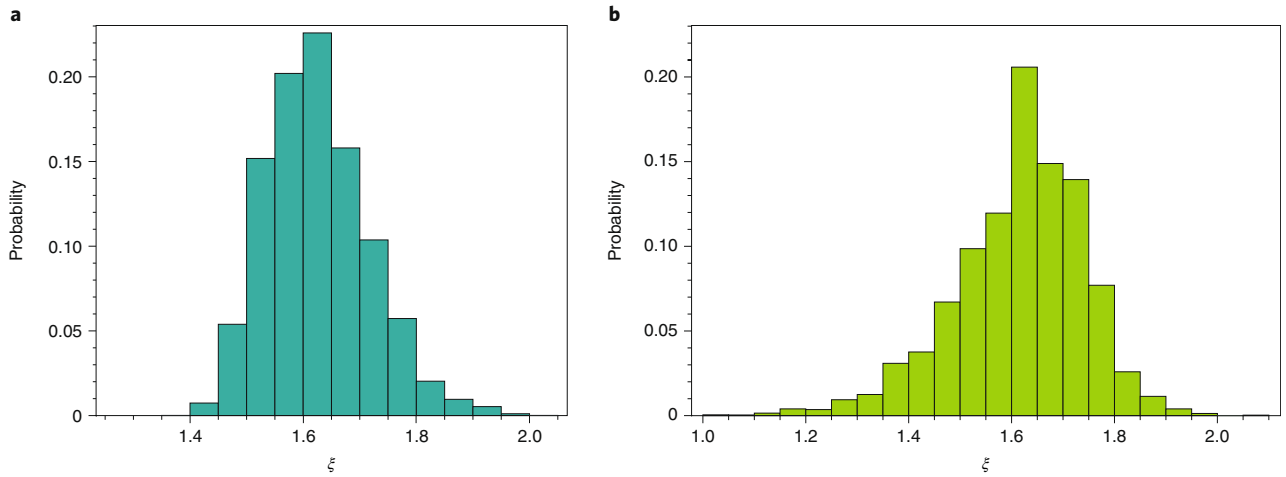


**Fig. 4 | Historical imprecision and missing observations. a**, Values of the shape parameter $\xi$ over 10,000 distorted copies of the dual versions of the average deaths in Table 1, allowing for a random variation of ±20% for each single observation, to account for the likely imprecisions in the historical data. The $\xi$ parameter consistently indicates an apparently infinite-mean phenomenon. **b**, Values of the shape parameter $\xi$ over 10,000 jack-knifed versions of the dual versions of the actual average numbers in Table 1, when allowing at least 1% and up to about 10% of the observations to be missing or redundant. The $\xi$ parameter consistently indicates an apparently infinite-mean phenomenon.

of $Y$, and $\varphi$ is monotonic. We call the distributions of $Y$ and $Z$ the real and the dual distribution, respectively. It can be verified that for values smaller than $M << H$, $Y$ and $Z$ are, in practice, indistinguishable (and so are their quantiles[8]).

As described in refs. [7,8], we take the observations in the 'Average estimate' column of Table 1, our $Y$ values, and transform them into their dual $Z$ values. We then study the unbounded duals using EVT, to find out that the naive observation of infinite moments can make sense in such a framework (but not for the bounded world population). Finally, by reverting to the real distribution, we compute the so-called shadow mean[8] of pandemics, equal to

$$E[Y] = (H-L)e^{\frac{1}{\xi}\frac{\sigma}{H}}\left(\frac{\sigma}{H\xi}\right)^{\frac{1}{\xi}}\Gamma\left(1-\frac{1}{\xi},\frac{\sigma}{H\xi}\right)+L, \qquad (2)$$

where $\Gamma$ is the Gamma function.

Notice that the random quantity $Y$ is defined above $L$, therefore its expectation corresponds to a tail expectation with respect to the random variable $Z$, an 'expected shortfall' in financial jargon[7]. All moments of the random variable $Y$ are called shadow moments in ref. [8], as they are not immediately visible from the data, but from plug-in estimation.

## The dual tail via EVT and the shadow mean

Take the dual random variable $Z$ whose distribution function $G$ is unknown. Given a finite threshold $u$, we can define the exceedance distribution of $Z$ as

$$G_u(z) = P(Z \le z | Z > u) = \frac{G(z) - G(u)}{1 - G(u)} \qquad (3)$$

for $z \ge u$.

For a large class of distributions $G$, and high thresholds $u \to \infty$, $G_u$ can be approximated by a three-parameter generalized Pareto distribution (GPD)[2], i.e.

$$G_u(z) \approx \mathrm{GPD}(z;\xi,\beta,u) = \begin{cases} 1-\left(1+\xi\frac{z-u}{\beta}\right)^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1-e^{-\frac{z-u}{\beta}} & \xi = 0 \end{cases}, \qquad (4)$$

where $z \ge u$ for $\xi \ge 0$, $u \le z \le u - \beta/\xi$ for $\xi < 0$, $u \in \mathbb{R}$, $\xi \in \mathbb{R}$ and $\beta > 0$.

Let us just consider $\xi > 0$, which is the case for fat tails[3]. From equation (3), we see that $G(z) = (1 - G(u))G_u(z) + G(u)$, hence we obtain

$$\begin{aligned} G(z) &\approx (1-G(u))\mathrm{GPD}(z;\xi,\beta,u) + G(u) \\ &= 1 - \bar{G}(u)\left(1+\xi\frac{z-u}{\beta}\right)^{-1/\xi}, \end{aligned}$$

with $\bar{G}(x) = 1 - G(x)$. The tail of $Z$ is therefore

$$\bar{G}(z) = \bar{G}(u)\left(1+\xi\frac{z-u}{\beta}\right)^{-\frac{1}{\xi}}. \qquad (5)$$

Equation (5) is called the tail estimator of $G(z)$ for $z \ge u$. Given that $G$ is, in principle, unknown, one usually substitutes $G(u)$ with its empirical estimator $n_u/n$, where $n$ is the total number of observations in the sample, and $n_u$ is the number of exceedances above $u$.

Equation (5) then becomes

$$\bar{G}(z) = \frac{n_u}{n}\left(1+\xi\frac{z-u}{\beta}\right)^{-1/\xi} \approx 1 - \mathrm{GPD}(z^*;\xi,\sigma,u), \qquad (6)$$

where $\sigma = \beta\left(\frac{n_u}{n}\right)^{\xi}$, $\mu = u - \frac{\beta}{\xi}\left(1-\left(\frac{n_u}{n}\right)^{\xi}\right)$, and $z^* \ge \mu$ is an auxiliary variable. Both $\sigma$ and $\mu$ can be estimated semi-parametrically, starting from the estimates of $\xi$ and $\beta$ in equation (4). Since we are considering the case $\xi > 0$, the preferred estimation method is maximum likelihood[2,3]. For both the exceedances distribution and the recovered tail, the parameter $\xi$ is the same, and it also coincides with the tail parameter we have used to define fat tails.

One can thus study the tail of $Z$ without worrying too much about the rest of the distribution, that is, the part below $u$. All in all, the most destructive risks come from the right tail, and not from the first quantiles or even the bulk of the distribution. The identification of the correct u is a relevant question in extreme value statistics[2,3]. One can rely on heuristic graphical tools[6], like the Zipf

plot and the meplot we plotted above, or on statistical tests for extreme value conditions[14] and GPD goodness-of-fit[21].

What is important to stress—once again—is that the GPD fit needs to be performed on the dual quantities, to be statistically and epistemologically correct. One could in fact work with the raw observation directly, without the log-transformation of equation (1), surely ending up with $\xi > 1$, in line with Figs. 1 and 2d. But such an approach would be incorrect because only the dual observations are actually unbounded.

Working with the dual observations, we find out that the best GPD fit threshold is around 200,000 victims, with 34.7% of the observations lying above this value. For the GPD parameters, we estimate $\xi = 1.62$ (standard error 0.52), and $\beta = 1.1747 \times 10^6$ (standard error $5.365 \times 10^5$). As expected, $\xi > 1$ once again supporting the idea of an infinite first moment. Visual inspections and statistical tests[14,21] support the goodness-of-fit for the exceedance distribution and the tail.

Looking at the standard error of $\xi$, one could also argue that, with more data from the upper tail, the first moment could possibly become finite. Yet there is no doubt about the non-existence of the second moment, and thus about the unreliability of the sample mean[5], which remains too volatile to be safely used. Pandemic fatalities would still be an extremely erratic phenomenon, with substantial tail risk. In any case, Figs 1 and 2d lead us to consider the first moment as infinite, and not to trust sample averages.

Given $\xi$ and $\beta$, we can use equations (2) and (6) to compute the shadow mean of the numbers of victims in pandemics. For actual data we get a shadow mean of 20.1 million, which is definitely larger (almost 1.5 times) than the corresponding sample tail mean of 13.9 million (this is the mean of all the actual numbers above the 200,000 threshold). Combining the shadow mean with the sample mean below the 200,000 threshold, we get an overall mean of 7 million instead of the value 4.9 million we have computed initially. It is therefore important to stress that a naive use of the sample mean would be misleading, inducing a substantial underestimation of risk. Under fat tails, averages are very often tricky objects.

Other sample quantities and moments could be computed via the dual approach[8], as the large (yet finite) upper bound guarantees their existence. Our argument is that it is these quantities, rather than the naive sample equivalents, or the reassuring point estimates of models underestimating tail risk, that should be the starting point of policy discussions.

## Data reliability issues

Clearly, estimates of the number of victims in past pandemics are not at all unique and precise. Figures are very often anecdotal, based on citations and vague reports, and usually dependent on the source of the estimate. In Table 1, it is evident that some events vary considerably in estimates.

Natural questions thus arise: are the tail risk estimates we have presented robust? What happens if some of the casualties estimates change? What is the impact of ignoring some events in our collection? The use of extreme value statistics in studying tail risk already guarantees the robustness of our estimates to changes in the underlying data, when these lie below the threshold $u$. However, to verify robustness more rigorously and thoroughly, we have decided to stress the data, to study how the tail may potentially vary.

First of all, we have generated 10,000 distorted copies of our dual data. Each copy contains exactly the same number of observations as per Table 1, but every data point has been allowed to vary between 80% and 120% of its recorded value before imposing the log transformation of equation (1). In other words, each of the 10,000 new samples contains 72 observations, and each observation is a (dual) perturbation (20%) of the corresponding observation in Table 1.

Figure 4a contains the histogram of the $\xi$ parameter over the 10,000 distorted copies of the dual numbers. The values are always above 1, indicating an apparently infinite mean, and the average value is 1.62 (standard deviation 0.10), in line with our previous findings. Our tail estimates are thus robust to imprecise observations. We find consistent results also for the $\beta$ parameter.

It also true that our data set is likely to be incomplete, not containing all epidemics and pandemics with more than 1,000 victims, or that some of the events we have collected are too biased to be reliable and should be discarded anyway. To account for this, we have once again generated 10,000 copies of our sample using a jack-knife resampling technique. Each new dual sample is obtained by removing from 1 to 7 observations at random, so that one sample could, for example, not contain the Spanish flu, while another could ignore the Yellow Fever and AIDS. In Fig. 4b we show the impact of such a procedure on the $\xi$ parameter. Once again, it is clear that it is robust to these significant changes in the underlying data set.

In conclusion, the central message based on this work is that pandemics are a fat-tailed phenomenon, with an extremely large tail risk and potentially destructive consequences. These should not be downplayed in any serious policy discussion.

Received: 1 April 2020; Accepted: 28 April 2020;
Published online: 25 May 2020

## References

1. Albert, R. & Barabasi, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74**, 47 (2002).
2. de Haan, L. & Ferreira, A. *Extreme Value Theory: An Introduction* (Springer, 2006).
3. Embrechts, P., Klüppelberg, C. & Mikosch, T. *Modelling Extremal Events* (Springer, 2003).
4. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-law distributions in empirical data. *SIAM Rev.* **51**, 661–703 (2009).
5. Taleb, N. N. *Statistical Consequences of Fat Tails* (STEM Academic Press, 2020).
6. Cirillo, P. Are your data really Pareto distributed? *Physica A* **392**, 5947–5962 (2013).
7. Cirillo, P. & Taleb, N. N. On the statistical properties and tail risk of violent conflicts. *Physica A* **452**, 29–45 (2016).
8. Cirillo, P. & Taleb, N. N. Expected shortfall estimation for apparently in finite-mean models of operational risk. *Quant. Finance* **16**, 1485–1494 (2016).
9. Nešlehová, J., Embrechts, P. & Chavez-Demoulin, V. Infinite-mean models and the LDA for operational risk. *J. Op. Risk* **1**, 3–25 (2006).
10. Taleb, N. N. & Cirillo, P. The decline of violent conflict: what do the data really say? In *Nobel Symposium Proc.* (eds Toje, A. & Bård, N. V. S.) 57–85 (Norwegian Nobel Institute, 2019).
11. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653 (2000).
12. Ferguson, N. et al. *Report 9: Impact of nonpharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand* (2020); https://www.imperial.ac.uk/media/imperial-college/medicine/mrc-gida/2020-03-16-COVID19-Report-9.pdf
13. Donnat, C. & Holmes, S. Modeling the heterogeneity in COVID-19's reproductive number and its impact on predictive scenarios. Preprint at https://arxiv.org/abs/2004.05272 (2020).
14. Falk, M., Hüsler, J. & Reiss, R.-D. *Laws of small numbers: extremes and rare events* (Birkhäuser, 2004).
15. Norman, J., Bar-Yam, Y. & Taleb, N. N. *Systemic risk of pandemic via novel pathogens - coronavirus: a note* (New England Complex Systems Institute, 2020).
16. Seybolt, T. B., Aronson, J. D. & Fischho, B. (eds) *Counting Civilian Casualties, An Introduction to Recording and Estimating Nonmilitary Deaths in Conflict* (Oxford Univ. Press, 2013).
17. Goldewijk, K., Beusen, K. & Janssen, P. Long term dynamic modeling of global population and built-up area in a spatially explicit way, hyde 3.1. *The Holocene* **20**, 565–573 (2010).
18. Klein Goldewijk, K. & van Drecht, G. HYDE 3.1: Current and historical population and land cover. In *Integrated modelling of global environmental change. An overview of IMAGE 2.4* (eds Bouwman, A. F., Kram, T. & Klein Goldewijk, K.) 93–112 (Netherlands Environmental Assessment Agency, 2006).
19. *2015 Revision of World Population Prospects* (United Nations, 2015)
20. World Health Organization. *Coronavirus Disease (COVID-19) Pandemic* https://www.who.int/emergencies/diseases/novel-coronavirus-2019 (WHO, accessed 13 April 2020).

21. Arshad, M., Rasool, M. T. & Ahmad, M. I. Anderson Darling and modified Anderson Darling tests for generalized Pareto distribution. *J. Appl. Sci.* **3**, 85–88 (2003).
22. List of epidemics. *Wikipedia* https://en.wikipedia.org/wiki/List_of_epidemics (accessed 30 March 2020).
23. Plague in the ancient & medieval world. *Ancient History Encyclopedia* https://www.ancient.eu/article/1528/plague-in-the-ancient--medieval-world/ (accessed 30 March 2020).
24. List of epidemics compared to coronavirus. *ListFist.com* https://listfist.com/list-of-epidemics-compared-to-coronavirus-covid-19 (accessed 30 March 2020).
25. Scasciamacchia, S. et al. Plague epidemic in the Kingdom of Naples, 1656-1658. *Emerg. Infect. Dis.* **18**, 186–188 (2012).
26. Great Northern War plague outbreak. *Wikipedia* https://en.wikipedia.org/wiki/Great_Northern_War_plague_outbreak (accessed 30 March 2020).
27. Visualizing the history of pandemics. *Visual Capitalist* https://www.visualcapitalist.com/history-of-pandemics-deadliest/ (accessed 30 March 2020).

## Competing interests
The authors declare no competing interests.

## Additional information
**Correspondence** should be addressed to P.C. or N.N.T.