Tail risks and infectious disease: Influenza mortality in the U.S., 1900–2018☆

Michele Campolieti

Department of Management, University of Toronto Scarborough, 1095 Military Trail, Toronto, Ontario, M1C 1A4, Canada

ARTICLE INFO

ABSTRACT

I use extreme values theory and data on influenza mortality from the U.S. for 1900 to 2018 to estimate the tail risks of mortality. I find that the distribution for influenza mortality rates is heavy-tailed, which suggests that the tails of the mortality distribution are more informative than the events of high frequency (i.e., years of low mortality). I also discuss the implications of my estimates for risk management and pandemic planning.

© 2021 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Recently, Cirillo and Taleb (2020) studied mortality in major pandemics, from ancient times to present, and focused their analysis on estimating heavy-tailed distributions using extreme values theory. Unlike current methods used for forecasting and modelling disease outbreaks (e.g., those described in the overviews provided in Hethcote, 2000; and, Unkel et al., 2012), extreme values theory is concerned with the tails of the distribution, not the mean or the bulk of the observations. As noted by Cirillo and Taleb (2020), extremes value theory emphasizes that there is more information in the tails of the distribution than in the outcomes that are more frequently observed. This focus means that extreme values theory is more suitable for planning for scenarios with extreme mortality, such as pandemics where there are new strains of viruses to which there is no immunity in the population. Using extremes value theory to plan for pandemics and for risk management has many parallels in emergency planning more broadly (e.g., flooding, de Haan & Ferreira, 2006; earthquakes, Pisarenko & Sornette, 2003). While extremes value theory is frequently used in planning and preparing for climatic disasters, it is still relatively new in the study of disease outbreaks. For example, there have been some analyses of weekly influenza mortality in China and France (Chen et al., 2015; Thomas et al., 2016), influenza mortality of those aged 65 and older in selected U.S. states (Lee & Wackernagel, 2007), Salmonella outbreaks (Guillou et al., 2014), Dengue Fever outbreaks (Lim et al., 2020), hospitalizations due to cardiovascular disease (Chiu et al., 2018) as well as studying Corona virus super-spreaders (Wong & Collins, 2020).
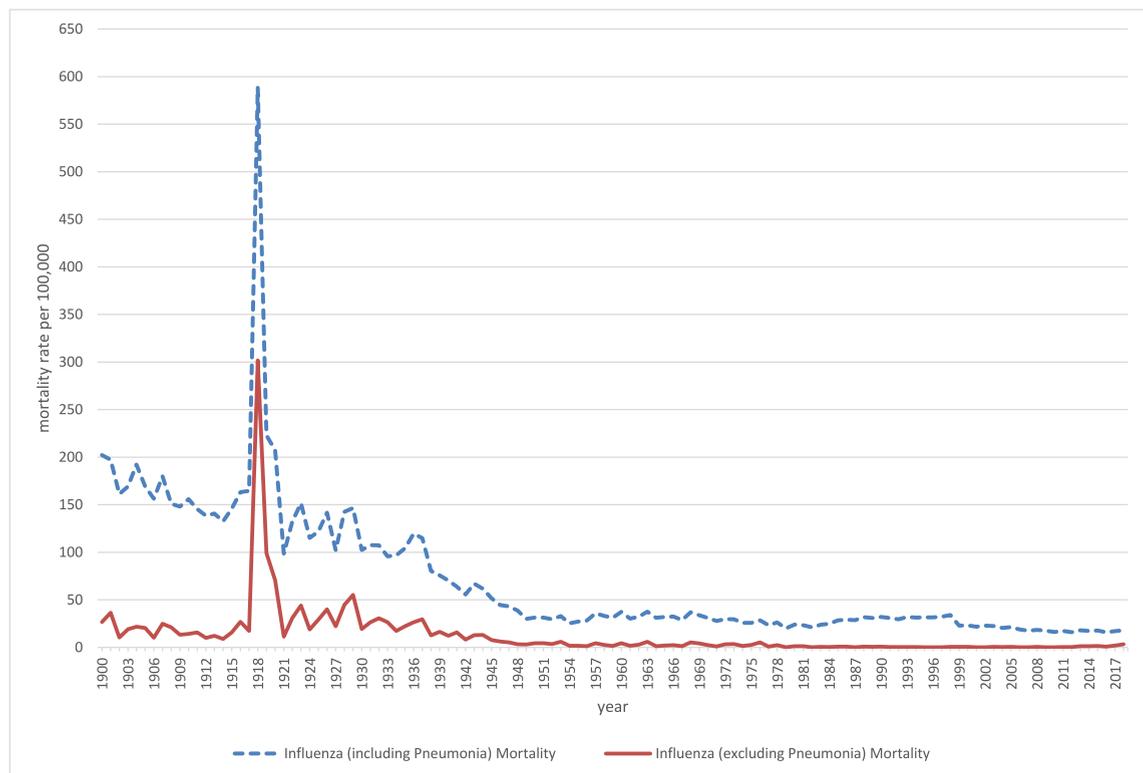

Fig. 1. Influenza mortality rates per 100,000, United States, 1900—2018

In this paper, I study the tail risks of infectious disease mortality. My analysis focuses on influenza and I estimate the tail risks using both pandemic and non-pandemic mortality, with U.S. data from 1900 to 2018. I use a maximum likelihood estimator that focuses on the years with elevated mortality, i.e., those that exceed a threshold, which includes both pandemic and epidemic outbreaks of influenza. I focus on the mortality rate, measured as the crude mortality rate per 100,000, and estimate a Generalized Pareto distribution. My estimates indicate that the distributions I estimate for influenza mortality are heavy-tailed. I also provide some assessment of the tail risks for elevated mortality for influenza, which is useful for modelling and planning for future outbreaks with elevated mortality.

2. Material and methods

I obtain the data I use from publications collected by the United States Center for Disease Control (CDC). My study period includes 1900 to 2018. I use the crude mortality rates for the years from 1900 to 1998 (per 100,000 of the population) for the United States reported in CDC (undated), except for 1958 and 1959 which were obtained from Grove and Hetzel (1968). Mortality rate data for 1999 to 2018 were obtained from the CDC website data retrieval portal on leading causes of death. Causes of death in these publications are listed by ICD code (various versions). The data for 1900 to 1932 are for the death registration area. The death registration area did not initially include all U.S. states, but the number of states included did increase over time.[1] Beginning in 1933 all states were part of the U.S. vital statistics system.

I consider two measures of influenza of mortality. First, I measure influenza mortality as the deaths due to influenza and pneumonia, i.e., influenza (including pneumonia) mortality, as is the convention in most of the literature (e.g., among others, Choi & Thacker, 1981; Noymer, 2008; Thompson et al., 2009). Many individuals who have influenza can also develop pneumonia and in the more serious cases death results from the pneumonia, which will be listed as the cause of death on the death certificate. Second, I follow Doshi (2008) who considered influenza mortality without including deaths from pneumonia and argued that it was a reliable measure of mortality.[2]

---

[1] Details on the death registration areas during this period can be found in Table 1.01 in Appendix II and the accompanying discussion in Centre for Disease Control (1997).

[2] Mortality rates for influenza (excluding pneumonia) were obtained from the following CDC publications: 1900—1940 are from Linder and Grove (1943), 1940—1960 are from Grove and Hetzel (1968); data for 1960—1998 were taken from the unpublished tables available on https://www.cdc.gov/nchs/nvss/mortality/hist290a.htm; 1999—2018, were obtained from CDC website data retrieval portal.

Influenza pandemics occur when there is a novel strain of the Influenza A virus to which populations have no natural immunity. During the period I consider there were four influenza pandemics: 1) the Spanish Flu in 1918−1919 (which was likely a H1N1 influenza virus, Taubenberger & Morens, 2006; Jester et al., 2018); 2) the Asian Flu in 1957−1959 (H2N2); the Hong Kong Flu in 1968−1969 (H3N2); and, 4) the H1N1 Flu in 2009. There was also an influenza pandemic in 1899−1900, but I only have the mortality data for 1900. There were also three notable epidemic years during my study period (Kilbourne, 2006): 1947; 1976, which was an aborted and potential pandemic influenza strain; and, 1977, which was a pandemic among the young (Russian Flu).

I plot the annual mortality rates for influenza (including pneumonia) in Fig. 1. As can be seen in the figure, influenza (including pneumonia) mortality during the Spanish Flu pandemic was quite elevated as it reached 589.5 per 100,000. However, there are quite a few years where the mortality rates exceed 100 per 100,000. Also notable is that the mortality from the pandemics in 1957−1958 and 1968−1969 can be lower than that in non-pandemic years. This was also recognized in Doshi (2008), who surmised that influenza pandemics are not as virulent as they were earlier in the 20th century, as the mortality rates are lower. However, this sort of observation neglects the tail risks inherent in the 1918−1919 Spanish Flu pandemic and the epidemics during the first few decades of the 20th century. I also present the mortality rates for influenza (excluding pneumonia) in Fig. 1. These mortality rates are lower than those that include pneumonia, but show the same sorts of spikes in mortality, e.g., the 1918−1919 pandemic and the large mortality during the epidemics in the 1930s. One difference between the mortality rates plotted in Fig. 1 is that the mortality rates for influenza (including pneumonia) show a steady decline during the first half of the 20th century before leveling off in the 1950s. In contrast, the mortality rates for influenza excluding pneumonia do not have a downward trend during the first half of the 20th century, but do level off at a lower level after the 1950s like the influenza mortality rates that include pneumonia in Fig. 1. Between 1900 and 2018 the median mortality rates for influenza (including pneumonia) and influenza (excluding pneumonia) were 32.3 and 3.4 per 100,000, respectively.

Let $x_1, \ldots, x_N$ be independently and randomly distributed random variables from an unknown distribution function F(x). As shown in Coles (2001), the distribution of values of x that exceed a threshold u is defined by

$$F_u(\varepsilon) = Prob[x > u + \varepsilon | x > u] = \frac{F(\varepsilon + u) - F(u)}{1 - F(u)}. \tag{1}$$

If the threshold level u is sufficiently large than the distribution $F_u(\varepsilon)$ can be approximated by a Generalized Pareto distribution, which has cumulative distribution function

$$G(x) = 1 - \left[1 + \xi\left(\frac{x - u}{\sigma_u}\right)\right]^{-\frac{1}{\xi}}, \tag{2}$$

where u is a threshold, $x > u$, σ_u is a scale parameter, which depends on the threshold u, and is such that $\sigma_u > 0$ and ξ is a shape parameter. In addition, the Generalized Pareto distribution has an infinite mean if $\xi > 1$ and infinite variance if $\xi > \frac{1}{2}$. The shape parameter, ξ, is also key for determining the behavior of the distribution. In particular, when $\xi > 0$ the distribution is heavy tailed (i.e., a Pareto), a Beta distribution when $\xi < 0$, and an exponential distribution as $\xi \to 0$. I obtain my estimates using maximum likelihood.[3] The estimates of the parameters of the Generalized Pareto distribution can also be used to compute return levels, i.e., a level that is exceeded once every mth year, which provides some indication of the likelihood of elevated levels of influenza mortality in the future. The return level, denoted x_m, can be computed as

$$x_m = u + \frac{\sigma_u}{\xi}\left[(m\zeta_u)^\xi - 1\right], \tag{3}$$

where $\zeta_u = Prob(x > u)$ and the other notation is as defined as above.

3. Results and discussion

Prior to estimating the Generalized Pareto distribution, I must select a value for the threshold. The choice of the threshold involves a trade-off between bias and variance (Coles, 2001). Selecting too high a value for the threshold can create bias. Conversely, selecting too low a value for the threshold results in more variance (less efficiency). To select the threshold, I first inspected the mean residual life plot and plots of the scale and shape parameters from different choices of threshold levels (Coles, 2001). These visual diagnostics suggest a threshold between the values of 120−140 for influenza (including pneumonia) mortality seem reasonable. Using threshold values from this range of plausible threshold values, I estimated the Generalized Pareto distribution. I then compared the estimated empirical quantiles from these alternative threshold choices

---

[3] An alternative maximum likelihood estimator is based on the maximum value of an outcome measure during a particular period, e.g., rainfall, storm surge, which is referred to as the block maxima approach (Coles, 2001). This approach results in a Generalized Extreme Value family of distributions that nests the Frechet, Gumbel and Weibull types of distribution. However, since I am considering the cumulative mortality from influenza in a particular year, the block maxima approach is not the most appropriate way to model tail observations.
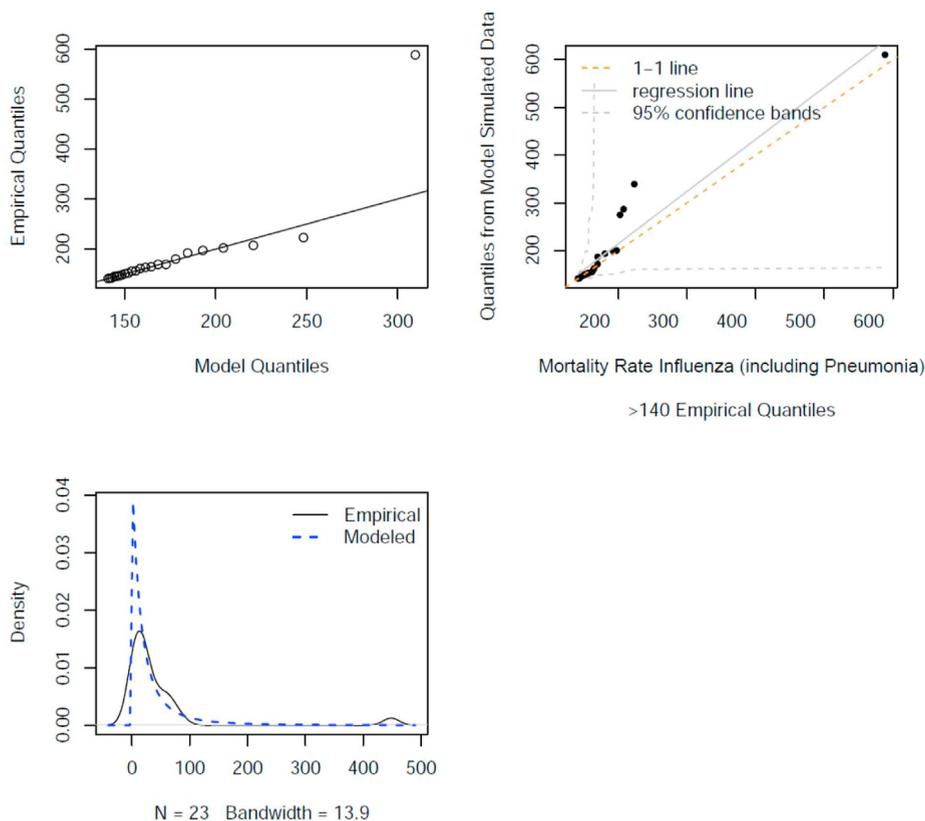
Fig. 2. Fit diagnostics, maximum likelihood estimates of influenza mortality (including pneumonia).

Table 1
Maximum likelihood estimates of generalized Pareto distribution, influenza mortality (including pneumonia).
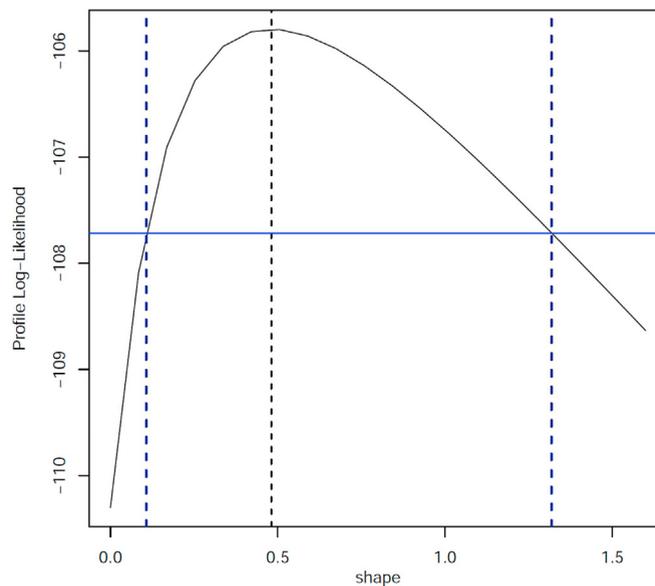
Parameter	Parameter Estimate	Confidence Interval
Scale parameter (σ_u)	22.589 (7.596)	[7.700,37.477]
Shape parameter (ξ)	0.481 (0.283)	[−0.073,1.035]
Log-likelihood	−105.80	
Number of observations exceeding threshold	23	
Number of observations	119	

Notes: Dependent variable is influenza (including pneumonia) mortality rate per 100,000. Standard errors in parentheses. Threshold value is mortality rate of 140 per 100,000. Estimates obtained using R package Extremes (Gilleland & Katz, 2016).

with the quantiles in the data and I the selected threshold value that produces empirical quantiles that best fit the quantiles in the data. Based on this approach, I selected the value of 140 for influenza (including pneumonia) mortality. I include a plot of the estimated quantiles against the empirical quantiles for this choice of threshold in Fig. 2.[4]

The estimates of the Generalized Pareto distribution with a threshold value of 140 are presented in Table 1. The shape parameter, ξ, has a point estimate of 0.481 and a standard error of 0.283. Consequently, the 95 percent confidence interval for ξ has a lower bound of −0.073 and an upper bound of 1.035. The negative values in the lower bound of the confidence interval mean that a Beta distribution, i.e., $\xi < 0$, and exponential distribution, $\xi = 0$, cannot be rejected. As discussed earlier $\xi > 0$ corresponds to a Pareto (i.e., heavy tailed) distribution. Since the upper bound of the confidence interval does exceed 1, a heavy tailed distribution with infinite mean and infinite variance is encompassed. I also follow the suggestion in Coles (2001) and compute a 95 percent confidence interval using the profile likelihood method, which can produce a confidence interval

---

[4] I also used maximum likelihood (Hill estimator) to estimate a power law distribution for influenza (including pneumonia) mortality with the approach in Clauset et al. (2007). Clauset et al. (2007) select the lower bound on the power law distribution by picking a value that maximizes the Kolmogorov-Smirnov test for the estimated distribution. Using this approach produces a lower bound of 138.4 for the power law distribution. This suggests that my choice of 140 as a threshold is a reasonable one.

Notes: Thicker vertical dashed lines denote lower and upper bounds of confidence interval.
Thinner vertical dashed line represents point estimate of shape parameter.

Fig. 3. Plot of confidence interval based on profile likelihood, shape parameter, ξ, influenza (including pneumonia) mortality. Notes: Thicker vertical dashed lines denote lower and upper bounds of confidence interval. Thinner vertical dashed line represents point estimate of shape parameter.

that has less parameter uncertainty. I plot the 95 percent confidence interval based on the profile likelihood method in Fig. 3. The profile likelihood confidence interval has a lower bound of 0.107 and upper bound of 1.318, and it is skewed to the right. This suggests that influenza (including pneumonia) mortality does have a heavy-tailed distribution and the confidence interval excludes the Beta or exponential distributions.

The estimates of the Generalized Pareto distribution for influenza mortality (excluding deaths from pneumonia) are presented in Table 2. I selected a threshold value of 20 for the analysis of these data, with the same approach I described earlier, as it produces the best fit of the model to the data (see Fig. 4).[5] The estimates of the shape parameter for the Generalized Pareto distribution fit to influenza (excluding pneumonia) mortality is 0.619 with a confidence interval that ranges between 0.012 and 1.226. The confidence interval excludes shape parameters of 0 or less, so there is strong evidence that the distribution is heavy tailed. In addition, the upper end of the confidence interval exceeds 1, which suggests that a distribution with an infinite mean and variance is encompassed. I also compute the 95 percent confidence interval using the profile likelihood method and present it in Fig. 5. The confidence interval based on the profile likelihood method has a lower bound of 0.19 and an upper bound of 1.51 as well as being skewed to the right. This confirms that the distribution of influenza (excluding pneumonia) mortality is heavy-tailed and the confidence interval also includes a distribution with an infinite mean.

I use the estimates of the Generalized Pareto distributions in Tables 1 and 2 to compute return levels, which are presented in Figs. 6 and 7. Note that the horizontal axis (i.e., the return period) in Figs. 6 and 7 uses a logarithmic scale. Consistent with a heavy tailed distribution, the return level plots trace out a concave pattern. The return levels provide some indication of the frequency of elevated mortality levels in the future. For example, for influenza (including pneumonia) mortality the return levels for the next 50 years range from 122.77 to 232.87 per 100,000. These return levels are 3.8–7.2 times greater than the median mortality for influenza (including pneumonia) during my study period. The return levels for influenza (excluding pneumonia) mortality indicate that mortality rates that range from 12.19 to 73.97 per 100,000, which are 3.8–21.8 times the median mortality rate for influenza (excluding pneumonia) during my study period. The probability of the mortality rates at the upper end of the return level (i.e., 232.87 and 73.97 per 100,000) for both measures I consider is about 0.10, which is relatively high when considering the mortality rates from influenza in the 2nd half of the 20th century. The estimates from the Generalized Pareto distribution place a much higher likelihood on elevated mortality events, like those that occur in pandemic or severe epidemic years. This difference results because the extremes value theory does not focus on the

---

[5] I also estimated a power law distribution for influenza (excluding pneumonia) with the Clauset et al. (2007) estimator, which selected a lower bound of 15.7 for the power law distribution.

Table 2

Estimates of generalized Pareto distribution, influenza mortality (excluding pneumonia).

Parameter	Parameter Estimate	Confidence Interval
Scale parameter (σ_u)	10.876 (3.811)	[3.467,18.345]
Shape parameter (ξ)	0.619 (0.309)	[0.012,1.226]
Log-likelihood	−92.123	
Number of observations exceeding threshold	23	
Number of observations	119	

Notes: Dependent variable is influenza (excluding pneumonia) mortality rate per 100,000. Standard errors in parentheses. Threshold value is mortality rate of 20 per 100,000. Estimates obtained using R package Extremes (Gilleland & Katz, 2016).
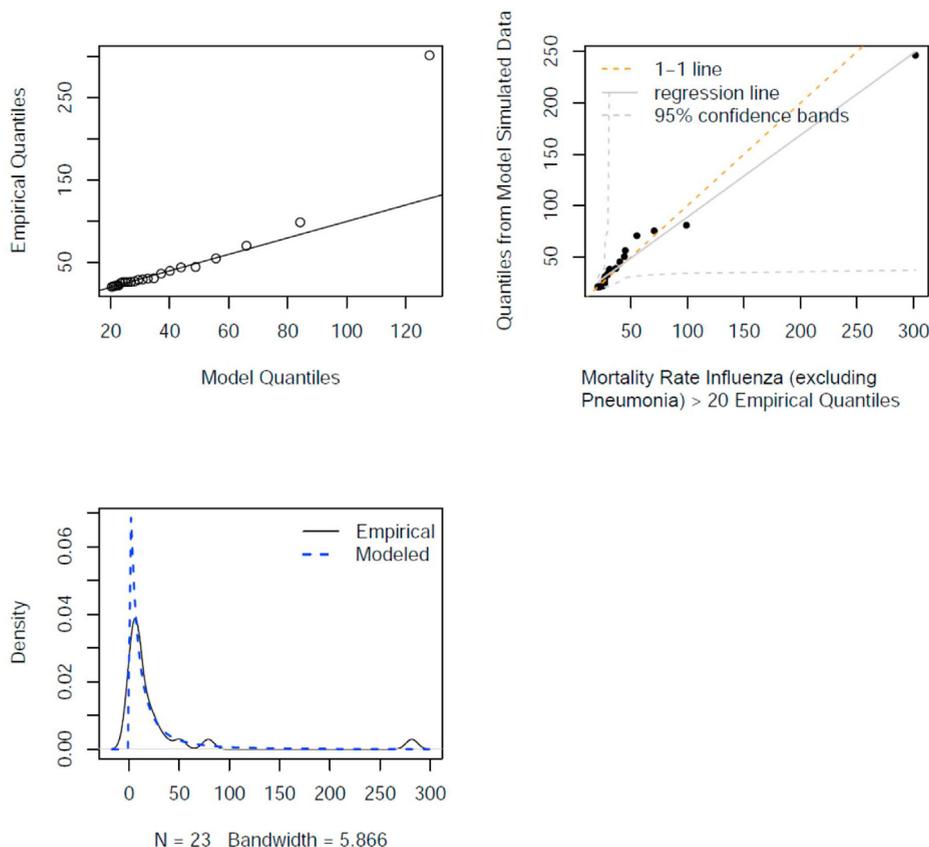

Fig. 4. Fit diagnostics, maximum likelihood estimates of influenza mortality (excluding pneumonia).

numerous years of low mortality for most of the last the last century, but rather the years of elevated mortality that exceed a threshold.

4. Conclusions

I used influenza mortality data to estimate the tail risks of mortality with U.S. data from 1900 to 1918. The estimates for both measures of influenza mortality I consider indicate that the data are consistent with a heavy-tailed distribution, although the evidence for a heavy-tailed distribution is much stronger when influenza mortality is measured excluding pneumonia. One limitation of my study is that the number of observations is not large, which can increase the variability in the estimates.

My estimates also imply that during a 50-year period influenza (including pneumonia) mortality rates would range from 3.8 to 7.2 times the median influenza (including pneumonia) mortality rate for the 1900 to 2018 period. For influenza (excluding pneumonia) mortality my estimates indicate a mortality rate which is 3.8–21.8 times the median mortality for the 1900–2018 period. My estimates based on extreme values theory indicate that these elevated mortality levels are much more
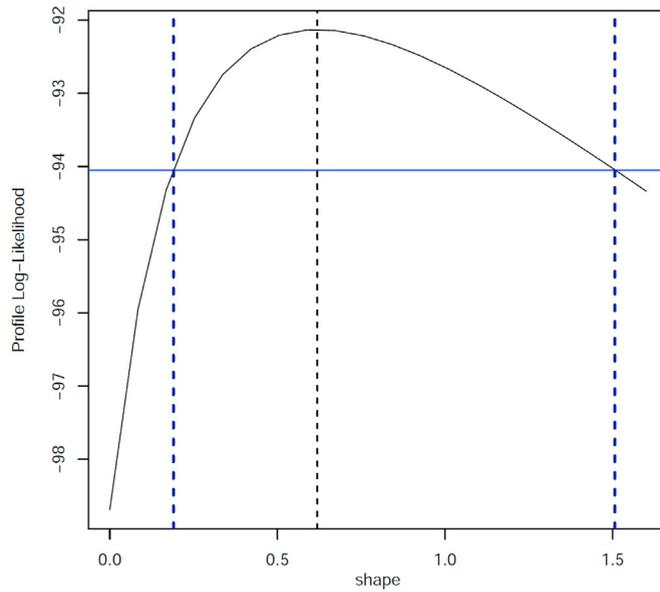
Notes: Thicker vertical dashed lines denote lower and upper bounds of confidence interval.
Thinner vertical dashed line represents point estimate of shape parameter.

Fig. 5. Plot of confidence interval based on profile likelihood, shape parameter, ξ, influenza (excluding pneumonia) mortality. Notes: Thicker vertical dashed lines denote lower and upper bounds of confidence interval. Thinner vertical dashed line represents point estimate of shape parameter.

likely than mortality rates from the 2nd half of the 20th century would indicate. This finding is somewhat striking as some have noted that planning and preparations for severe influenza pandemics is inadequate (Horimoto & Kawaoka, 2005; Jester et al., 2018). One implication of my findings is that extreme values theory can better assess the probability of future pandemics or epidemics with elevated mortality rates. For example, I find a mortality rate for influenza (including pneumonia) of 232.87 per 100,000 has a probability of about 0.10. In contrast, pandemic preparation protocols used by most countries have risk
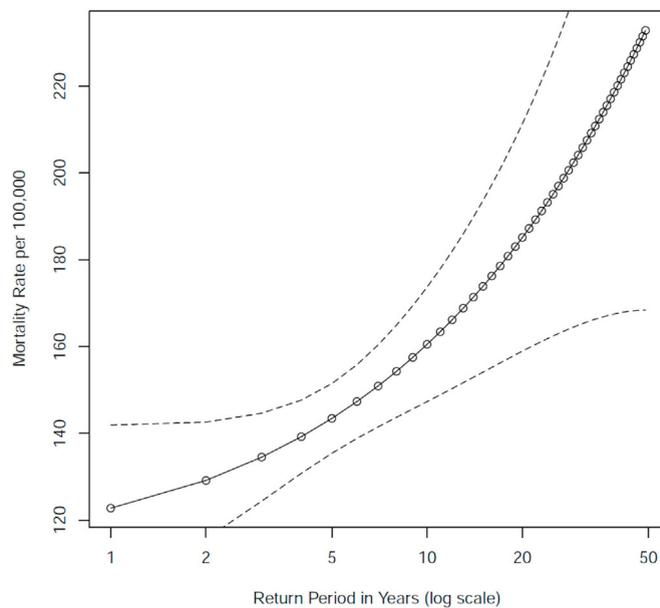
Notes: Dashed line is 95 percent confidence interval. Return period uses a log scale.

Fig. 6. Return level plot, influenza (including pneumonia) mortality. Notes: Dashed line is 95 percent confidence interval. Return period uses a log scale.
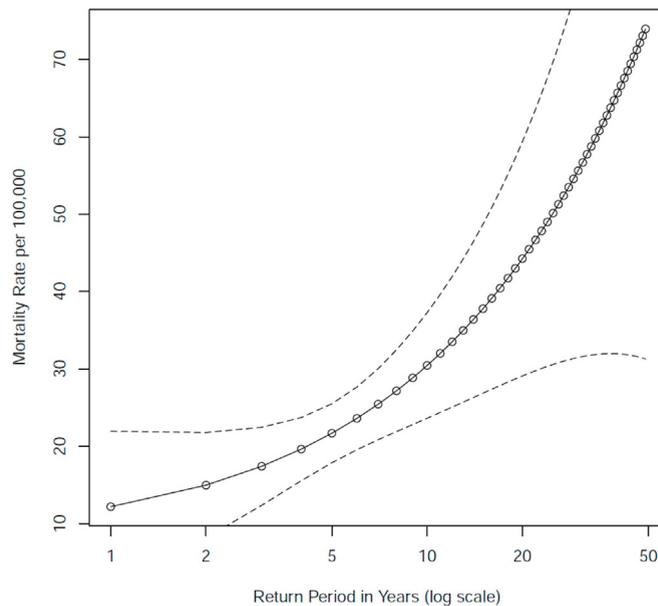
Notes: Dashed line is 95 percent confidence interval. Return period uses a log scale.

Fig. 7. Return level plot, influenza (excluding pneumonia) mortality. Notes: Dashed line is 95 percent confidence interval. Return period uses a log scale.

assessment tools that are more subjective in nature. For example, the CDC assesses ten risk elements for an influenza pandemic as low, moderate and high and uses a scoring system to determine whether the risk of a pandemic is low, moderate or high. However, the limitation of this approach is that it does not consider the tail risks of elevated mortality, like extremes value theory, and could underestimate the probability of an outbreak with high mortality. Greater awareness of the tail risks of elevated mortality might make public health agencies and governments reconsider their pandemic planning and help them better plan and prepare for pandemics (e.g., preparing suitable stockpiles of personal protective equipment or planning for surges in cases that require hospitalization).

Funding

Declaration of competing interest

References

Center for Disease Control (undated). Leading Causes of Death, 1900-1998. Obtained at https://www.cdc.gov/nchs/nvss/mortality_historical_data.htm (last accessed December 9, 2020).

Centre for Disease Control. (1997). *U.S. Vital statistics system: Major activities and developments* (pp. 1950—1995).

Chen, J., Lei, X., Zhang, L., & Peng, B. (2015). Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejiang China. *PloS One, 10,* Article e0118521. https://doi.org/10.1371/journal.pone.0118521

Chiu, Y., Chebana, F., Abdous, B., & Belanger, D. (2018). Mortality and morbidity peaks modeling: An extreme value theory approach. *Statistical Methods in Medical Research, 27,* 1498—1512.

Choi, K., & Thacker, S. B. (1981). An evaluation of influenza mortality surveillance, 1962-1979. *American Journal of Epidemiology, 113,* 215—226.

Cirillo, P., & Taleb, N. N. (2020). Tail risk of contagious disease. *Nature Physics, 16,* 606—613.

Clauset, A., Young, M., & Gleditsch, K. S. (2007). On the frequency of severe terrorist attacks. *Journal of Conflict Resolution, 51,* 58—87.

Coles, S. (2001). *An introduction to statistical modelling of extreme values.* London: Springer-Verlag.

Doshi, P. (2008). Trends in recorded influenza mortality: United States, 1900-2004. *American Journal of Public Health, 98,* 939—945.

Gilleland, E., & Katz, R. W. (2016). extRemes 2.0: An extreme value analysis package in R. *Journal of Statistical Software, 72.* https://www.jstatsoft.org/article/view/v072i08.

Grove, R. D., & Hetzel, A. M. (1968). *Vital statistics rates in the United States 1940-1960.* Washington D.C: U.S. Department of Health, Education and Welfare.

Guillou, A., Kratz, M., & Le Strat, Y. (2014). An extreme value theory approach for the early detection of time clusters: A simulation-based assessment and an illustration to the surveillance of Salmonella. *Statistics in Medicine, 33*, 5015–5027.

de Haan, L., & Ferreira, A. (2006). *Extreme value theory: An introduction*. London: Springer-Verlag.

Hethcote, H. (2000). Mathematical models of infectious diseases. *SIAM Review, 42*, 599–653.

Horimoto, T., & Kawaoka, Y. (2005). Influenza: Lessons from past pandemics, warnings from current incidents. *Nature Reviews Microbiology, 3*, 591–600.

Jester, B., Uyeki, T., & Jernigan, D. (2016). Readiness for a responding to a severe pandemic 100 Years after 1918. *American Journal of Epidemiology, 187*, 2596–2602.

Kilbourne, E. (2006). Influenza pandemics of the 20th century. *Emerging Infectious Diseases, 12*, 9–14.

Lee, H. C., & Wackernagel, H. (2007). *Extreme values analyses of US P&I mortality data under consideration of demographic effects*. Fontainebleau, France: Centre de Géosciences. Ecole des Mines de Paris.

Lim, J. T., Dickens, B. S. L., & Cook, A. (2020). Modelling the epidemic extremes of Dengue transmission in Thailand. *Epidemics, 33*, 100402.

Linder, F. E., & Grove, R. D. (1943). *Vital statistics rates in the United States, 1900-1940*. Washington, DC: Department of Commerce, Bureau of the Census.

Noymer, A. (2008). Influenza analysis should include pneumonia. *American Journal of Public Health, 98*, 1927–1928.

Pisarenko, V. F., & Sornette, D. (2003). Characterization of frequency of extreme earthquake events by the generalized Pareto distribution. *Pure and Applied Geophysics, 160*, 2343–2364.

Taubenberger, J. K., & Morens, D. M. (2006). 1918 influenza: The mother of all pandemics. *Emerging Infectious Diseases, 12*, 15–22.

Thomas, M., Lemaitre, M., Wilson, M. L., Viboud, C., Yordanov, Y., Wackernagel, H., & Carrat, F. (2016). Applications of extreme value theory in public health. *PloS One, 11*, Article e1059312. https://doi.org/10.1371/journal.pone.0159312

Thompson, W. W., Moore, M. R., Weintraub, E., Cheng, P.-Y., Jin, X., Bridges, C. B., Bresee, J. S., & Shay, D. K. (2009). Estimating influenza-associated deaths in the United States. *American Journal of Public Health, 99*, S225–S230.

Unkel, S., Farrington, C. P., Garthwaite, P. H., Robertson, C., & Anders, N. (2012). Statistical methods for the prospective detection of infectious disease outbreaks: A review. *Journal of the Royal Statistical Assocation, Series A (Statistics in Society), 175*, 49–82.

Wong, F., & Collins, J. J. (2020). Evidence that coronavirus super-spreading is fat-tailed. *Proceedings of the National Academy of Sciences, 117*, 29417.