**Supplementary Material A**

**Data**

City-level data were compiled from Medical Officer of Health (MOH) reports. These reports were annual administrative documents covering a range of public health-related issues at the municipal level. The first reports begin in the mid-19th century, and coverage extends to most municipalities in the UK until the early 1970s. The reports used here have been digitised and can be viewed on the Wellcome Trust Collection website.27

We collected influenza mortality data between 1895 andfor eight large municipalities from across the UK: Belfast, Birmingham, Cardiff, Glasgow, London, Liverpool, Manchester and Sheffield. To do this, we searched relevant MOH reports for each municipality. Generally, annual mortality rates by cause of death are presented in tables within the report or its appendices (or could be computed as the ratio of the number of deaths to the population size). Despite changes to the taxonomy of many causes of death over time, *influenza* was reported in the MOH reports throughout the decades. During the war years, 1914-1919 and 1939-1945, some reports are missing, or the data provided is incomplete. In these cases, we recovered the missing information by assessing statistics for these years from later reports, where possible. The dataset collected from the MOH reports underlying this research paper is available onGitHub at <https://github.com/maxschr90/Schroeder-et-al.-2021--How-long-do-pandemics-last->.

The narrative provided in the MOH reports for London County Council confirms that the two decades after the three initial waves of the pandemic were characterised by several further large influenza outbreaks. In particular, we examined available London County Council MOH reports between 1920 and 1957 to identify discussions of increased prevalence of influenza in a given year. In each report, we performed a search for the phrase *influenza*. As public health officials were particularly attentive to influenza after the 1918-19 pandemic, virtually every annual report contains at least some discussion of the disease. Generally, the language of the reports is clear on whether a certain year is considered to have a notable outbreak of influenza. Some reports further included retrospective reflections on past influenza outbreaks. The reports identify 1922, 1924, 1927, 1929, 1933 and 1937 as years of heightened mortality from influenza relative to other years.

Data for the US are taken from the annual vital statistics reports compiled by the [National Center for Health Statistics](https://www.cdc.gov/nchs/index.htm). Specifically, we rely on the two special volumes16,17 covering the period 1900-1960. Both volumes are available on the CDC’s website.28 Data for England and Wales between 1838 and 1917 are taken from Langford (2002),18 Table 5, who compiles mortality rates from different sources.

To obtain COVID-19 mortality for the analysis in Figure 4, we combined the total number of UK COVID-19 deaths between 6th March 2020 and 6th of March 2021 (124,654).29 With mid-year population data from the ONS for 2020 (67,081,000).30 This implies a mortality rate of 1,858 per million. Pre-2020 deaths from influenza are taken from the 2018-19 total mortality figures for England & Wales from the ONS31 together with the mid-2018 population figures for England & Wales.32

**Supplementary Material B**

**Modelling of mortality risk dynamics**

The model in (1) – (2) assumes that mortality rates after the main waves of the pandemic are the outcomes of a sequence of bounded Pareto distributions, where the inverse of the tail index of these distributions decays exponentially over time. The parameters and scale the range of mortality rates that the model predicts. We fit the model to the data for each geographical unit and conditional on its experience of the pandemic. Hence, we choose the bounds to reflect the realised range of mortality rates for the geographical unit over the period modelled. Conditional on and , the two parameters and then determine the dynamics of mortality and disease outbreak risk via controlling the level and time evolution of probabilities of outcomes associated with the tail of the Pareto distributions.

We obtain and by maximising the likelihood function:

, (3)

given a sample of mortality rates . To maximise the log likelihood function, we use MATLAB’s fmincon routine, using a sequential quadratic programming algorithm. Derivatives are approximated by central numerical derivatives, and the relevant termination criteria are set to 1e-12. To account for potential nonconvexities and the presence of local maxima, we begin the maximisation from 1000 random seed values. The estimated parameters are summarised in Table S1.

Table S1: Maximum likelihood estimates of base model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **λ** |  | **dl** | **du** |
| **City average** | 0.151 | 76.867 | 24 | 3059 |
| Belfast | 0.119 | 24.859 | 22 | 3644 |
| Birmingham | 0.144 | 58.442 | 30 | 2497 |
| Cardiff | 0.184 | 360.214 | 13 | 2671 |
| Glasgow | 0.166 | 74.119 | 33 | 1812 |
| Liverpool | 0.166 | 126.778 | 23 | 1775 |
| London | 0.11 | 26.352 | 23 | 4458 |
| Manchester | 0.18 | 208.861 | 23 | 3067 |
| Sheffield | 0.169 | 202.69 | 12 | 4550 |
| **US** | 0.278 | 802.709 | 31 | 3018 |
| **England & Wales** | 0.111 | 7.684 | 113 | 574 |

The model in (1) – (2) allows mortality risk to vary with time step after the main pandemic waves, as it declines towards background mortality. The modelling is motivated by the empirical observations discussed under Figure 1: the period after the pandemic has higher variation in mortality, and thus a higher mortality risk, than background mortality (i.e. during the period before the pandemic and the period a few decades after the pandemic). These observations imply that mortality risk just after the main pandemic waves should start from a higher level, converging to background mortality. Under the assumption that risk is decreasing over time, the model in (1) – (2) describes the level and rate of decline as a function of the parameters and , which determine the sequence . Although the model constrains the inverse of the tail index to decay exponentially, it does not impose such a restriction on *probabilities* associated with the tails. As determined by (1), the probabilities of different mortality rates are a monotonic function of the tail index and thus decrease over time. However, the inverse of a convex function does not have a predetermined convexity. Hence, implies that tail probabilities decline over time, but not that their rate of decline is exponential. This feature is evident in the results.

The choice of the bounds, and scales the range of mortality rates. In the baseline application, for each city is set to the mortality rate of 1918 for that city, and to lowest mortality experienced in the long run in each city, or the average across cities for the model fitted to the average. For the US following the 1918-19 pandemic, and are the lowest post-1918 and the 1918 mortality rates observed in the US time series; for the 1890-91 post-pandemic mortality rate dynamics, these parameters refer to the lowest between 1891 and 1917, and the 1891 mortality rates in the respective time series for England and Wales. These parameters are also summarised in Table S1.

The advantage of letting and be specific to the geographical unit studied is that we allow for differences in the experience of the pandemic and in conditions that influence background infectious disease mortality to have a bearing on risk dynamics. By allowing for such latent variation (see, e.g. Table S2 in Supplementary Material C for variation in population and density across the cities in the UK), our finding of a general pattern of mortality and disease outbreak risk dynamics is stronger. In Figure S1-S2, we also present results for a model of the different cities in which we set and to be common across all cities, specifically to the lowest and highest mortality rate observed across them. As can be seen, the overall pattern is very similar across cities (compared with the results in Figures 2 and 3, even greater similarity is observed). However, the finding of generality as based on Figure S2 on its own is likely biased because of omitted variation between the cities, which can lead to excess similarity. In contrast, the results in Figures 2 and 3 reassure us that the general pattern remains similar even after controlling for variation between cities.

Figure S1: Model predicted influenza mortality rates following the 1918-19 pandemic (model with common bounds)

******

**Simulated median (solid black line), interquartile range (dark shading) and 80% prediction interval (light shading) are based on 1m random draws. Simulated outcomes are based on the model fitted using the average across cities. Upper and lower bounds are common and set to be the highest and lowest mortality rates within the sample of cities respectively. Data are overplotted in red. Data for UK cities are taken from Medical Officer for Health reports.**

Figure S2: Outbreak risk following the 1918-19 pandemic (model with common bounds)

******

**Outbreak probabilities computed from models fitted to each city. Upper and lower bounds are common and set to be the highest and lowest mortality rates within the sample of cities respectively.**

Technically, and can be estimated jointly with and using the time series following the main pandemic waves (i.e., without exploiting the information on the mortality effect of the main pandemic waves). In this case, (3) is maximised by choosing all four parameters, following the same optimisation methods as for the model in (1) – (2).Figures S3 and S4 Figures 2 and 3 under this approach. As can be seen, the model predicted mortality is close to the actual path of mortality, while disease outbreak probabilities for large outbreaks are effectively zero. Although fitting the model in (1) – (2) by estimating all four parameters from the time series of data following the main pandemic waves may be useful for some applications (e.g. if the interest is in summarising historical experience), conceptually, this is no longer a model of mortality and disease-outbreak *risk* dynamics, i.e. of the time evolution of the probability of disease outbreaks. In this case, and are interpreted as parameters chosen to maximise the fit of the process to the data, and thus are chosen by the optimisation routine to be in effect the maximum and minimum mortality observed between 1920 and 1950. In turn, this rules out the possibility of disease outbreaks that are higher than those observed *ex post*, even if theoretically they could have happened. An analysis of risk dynamics must account for the possibility of higher mortality than that that was actually observed. Our modelling approach views bounds as *possible* even if unrealised mortality rates, given the experience of the pandemic, which determines the upper bound, and given the expectation about steady state mortality that is unaffected by the pandemic, which determines the lower bound.

Figure S3: Model predicted influenza mortality rates following the 1918-19 pandemic (model with estimated bounds)

******

**Simulated median (solid black line), interquartile range (dark shading) and 80% prediction interval (light shading) are based on 1m random draws. Simulated outcomes are based on the model fitted using the average across cities. Upper and lower bounds are fitted to the data. Data are overplotted in red. Data for UK cities are taken from Medical Officer for Health reports.**

Figure S4: Outbreak risk following the 1918-19 pandemic (model with estimated bounds)

******

**Outbreak probabilities computed from models fitted to each city. Upper and lower bounds are fitted to the data (1920-1950) for each city.**

We next compare the approach of specifying the bounds of the Pareto distribution conditionally on the mortality range relevant to the geographical unit to a more agnostic approach of setting and to be determined by theoretical upper and lower bounds of mortality rates. Figures S5 and S6 in reproduce Figures 2 and 3 under the agnostic modelling approach. As can be seen, the pattern in Figure S5 is qualitatively similar with that in Figure 2, but the model predicted mortality rates can be unrealistically high with significant probabilities. Moreover, while in Figure S6 we see that the probability of a disease outbreak remains high until the 1940s, the dynamic pattern is different from that in Figure 3, implying disease outbreak risk that is very high initially and declines more rapidly. Overall, agnostic modelling of and confirms high and persistent mortality risk after the main pandemic waves, as well as similarity across geographical units, but, compared with the post-pandemic experience, it exaggerates it relative to the actual experience. On the other hand, exploiting the information pertinent to the pandemic in parameterising and leads to more accurate mortality risk predictions.

Figure S5: Model predicted influenza mortality rates following the 1918-19 pandemic (model with theoretical bounds)

******

**Simulated median (solid black line), interquartile range (dark shading) and 80% prediction interval (light shading) are based on 1m random draws. Simulated outcomes are based on the model fitted using the average across cities. Upper and lower bounds are set to 1,000,000 and 1 respectively. Data are overplotted in red. Data for UK cities are taken from Medical Officer for Health reports.**

Figure S6: Outbreak risk following the 1918-19 pandemic (model with theoretical bounds)

******

**Outbreak probabilities computed from models fitted to each city. Upper and lower bounds are set to 1,000,000 and 1 respectively.**

Next, we discuss the assumption that mortality rates after the main pandemic waves are determined by bounded Pareto distributions. The bounded Pareto distribution has been used in modelling outbreak fatality risk21 and offers flexibility and tractability in capturing important properties of a distribution of mortality rates in the context of our analysis. A model of the dynamics of mortality risk after the main pandemic waves needs to capture mortality risk in a period spanning high mortality rates, immediately after the main waves of the pandemic, and low mortality rates a few decades after the pandemic when the effect of the pandemic has died out. Hence it is important that the underlying mortality distribution in any given year can simultaneously account for a fat tail, and thus a potentially higher probability of disease outbreaks, and ensure that the mass of the distribution remains at the lower end of mortality. Over time, the tail becomes less important, tending to background mortality, and the distribution converges to a very high concentration toward the lower bound of mortality rates. The bounded Pareto distribution has the flexibility to account in a tractable manner for both characteristics: it has a fat tail while the most likely outcomes remain near the lower bound of mortality; and, conditional on the bounds, a transition in the tail probabilities is identified via changes in one parameter.

To further illustrate key points of the relevance of the Pareto distribution, we describe and fit a model with an alternative one-parameter distribution that allows for high probabilities for outcomes associated with the tail, the one-parameter Weibull-type distribution.33 In this case, mortality rates are drawn from:

(4)

where , noting that the tail contracts as decreases. Assume that:

. (5)

Conditional on the time process in (5), and thus conditional on the sequence , is independently distributed over time following (4). The likelihood is given by:

,

for a sample of mortality rates . Figures S7 and S8 reproduce Figures 2 and 3 under this modelling. The results from this model also reveal that mortality risk remains high for a long period after the main pandemic waves, and that its dynamic pattern is similar across geographical units. However, as can be seen, the predicted probabilities for disease outbreaks are higher than those in Figure 3. This is an implication of the one-parameter Weibull form that delivers a fat tail by shifting the mass of the distribution away from lower numbers.33 Moreover, disease outbreak risk inherits a rate of rapid decline from the exponential decay of , the Weibull parameter that determines the thickness of the tail.

Figure S7: Model predicted influenza mortality rates following the 1918-19 pandemic (One-parameter Weibull model)

******

**Simulated median (solid black line), interquartile range (dark shading) and 80% prediction interval (light shading) are based on 1m random draws. Simulated outcomes are based on the model assuming a sequence of one-parameter Weibull distributions fitted using the average across cities (1920-1950). Data are overplotted in red. Data for UK cities are taken from Medical Officer for Health reports.**

Figure S8: Outbreak risk following the 1918-19 pandemic (One-parameter Weibull model)

******

**Outbreak probabilities computed from models fitted to each city assuming a sequence of one-parameter Weibull distributions.**

The analysis above illustrates the importance of the property of the Pareto distribution that it can accommodate a fat tail with the mass of the distribution near the lower bound of outcomes. Alternative distributions that can deliver a concentration at lower mortality levels while also allowing for high probabilities associated with tail outcomes require more parameters to be specified (e.g. lognormal, Gaussian mixture) and require more assumptions regarding the dynamic transition. In particular, the model must specify the dynamic evolution of two or more parameters and a means to identify the specific combination of the dynamic processes of these parameters that characterises the evolution of tail probabilities and of mortality risk more generally. Data availability restricts these options. Being a one-parameter distribution conditional on the bounds for mortality rates, the bounded Pareto offers a transparent way to model the post-pandemic dynamic evolution of mortality and disease outbreak risk.

**Mortality risk predictions following COVID-19 under model uncertainty**

We are interested in making predictions of probabilities of disease outbreaks at the point in time towards the end of the main waves of a pandemic, when there is information on the mortality of the main waves of the pandemic, and an expectation about the level of background mortality that should be reached after some time. The generality of the main qualitative characteristics of the dynamic risk patterns in Figures 2 and 3, across different geographical units with different experience of main wave and background, suggests that the model in (1) – (2) can be used for counterfactual analysis. Nevertheless, the specific quantitative predictions may depend on the specific sequence used, or equivalently on the specific parameters and . Inspection of Table S1 suggests that even though these parameters are relatively similar, they still vary across geographical units. This variation generates model uncertainty regarding the risk projections, because it creates uncertainty about which sequence of distributions in generate the data, which is different from the epidemiological uncertainty regarding mortality rates that is generated by a given sequence of distributions.

To account for such model uncertainty, we assume that the parameters and that are relevant to a new pandemic are drawn from the same distribution from which the parameters in Table S1 are drawn, and then perform a Monte Carlo analysis that provides a distribution of possible outcomes as a function of draws of and from that distribution. Using the parameter values of and in Table S1, we estimate the underlying distribution (see Figure S9) and approximate it as a joint lognormal:

We then draw one million pairs of and from the implied joint density and summarise relevant percentiles of the generated distribution of predicted probabilities of mortality rates in Figure 4.

Figure S9: Distribution of η0 and λ implied by the models fitted to previous pandemics

******

**Kernel density estimates of η0 and λ for different geographical regions.**

**Supplementary Material C**

**Additional Tables and Figures**

Table S2: Population and population density for UK cities

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Population | | | Population density | | |
|  | 1920s | 1930s | 1940s | 1920s | 1930s | 1940s |
| Belfast | 422,130 | 425,202 | 442,935 | - | - | - |
| Birmingham | 947,923 | 1,023,811 | 1,053,157 | 21 | 20 | 21 |
| Cardiff | 219,894 | 222,658 | 227,861 | 21 | 18 | 17 |
| Glasgow | 1,079,858 | 1,088,829 | 1,089,368 | 51 | 35 | 27 |
| Liverpool | 837,595 | 858,783 | 720,112 | 40 | 32 | - |
| London | 4,540,000 | 4,230,295 | 2,863,548 | - | - | - |
| Manchester | 749,970 | 746,974 | 641,040 | 35 | 29 | 24 |
| Sheffield | 519,224 | 517,967 | 492,093 | 17 | 14 | 13 |
| **Population and population density numbers refer to decadal averages compiled form available data in the MOH reports. Population density is measured in persons/acre.** | | | | | | |

Figure S10: Model predicted influenza mortality rates following the 1918-19 and 1890-91 pandemics (deaths per million)

**

**Simulated median (solid black line), interquartile range (dark shading) and 80% prediction interval (light shading) are based on 1m random draws. Simulated outcomes are based on the model fitted to each city, the US and England & Wales. Data for UK cities are taken from Medical Officer for Health reports. Data for the US is taken from Lindner and Grove (1943)**16 **and Grove and Hetzel (1968).**17 **Data for England & Wales are taken from Langford (2002),**18 **Table 5.**

Figure S11: Outbreak risk after the 1890-91 pandemic

******

**Outbreak probabilities computed from model fitted to data for England and Wales.**