Estimating the Reproductive Number of an Influenza Pandemic in Real-Time from Multiple Data Sources.

Carlo G. Camarda Jim Oeppen

Max Planck Institute for Demographic Research, Rostock, Germany

Short Abstract

An influenza epidemic is the outcome of an interaction between two populations – the virus and the human immune system – with very different demographic characteristics. Accurate real-time estimation of the parameters of this interaction, together with their confidence intervals, would be of enormous help to health planners. Conventional weekly reporting of influenza mortality and morbidity is being overtaken by real-time computerised medical record-keeping, rapid laboratory diagnosis, and indirect sources such as Google Flu trends. Our model aims to estimate the daily Reproductive Number of an epidemic in its growth phase (NRR in demographic terms) , while explicitly accounting for the mis-recording created by week-ends and public holidays. To evaluate the model we use data from two influenza pandemics: 1889-90 in Munich and 1918 in New York State. Treating these data as an *ex ante* estimation problem shows how well this key parameter might be estimated in a future pandemic.

Extended Abstract

Motivation.

The Reproductive Number of an epidemic (R_0) is the number of secondary cases produced by each primary case in a totally susceptible population. For an epidemic to die out, R_0 must fall below unity. Large values indicate epidemics that may not be susceptible to interventions designed to reduce transmissibility below the unit threshold. Demographers are already familiar with these concepts from population dynamics as the Reproductive Number is equivalent to the Net Reproduction Rate and the Serial Interval is equivalent to generation length.

In the early stages of an epidemic or pandemic, R_0 together with the Serial Interval between primary and secondary cases, is an important guide to the transmissibility of the disease and the possibility for intervention. Models that provide an early and accurate estimation of the reproductive number, and its confidence interval, are crucially important for health planners.

The Growth of Real-Time Influenza Data.

Infectious disease incidence data come from several sources. Sentinel systems recruit medical practitioners to report on patient cases. The details typically available for influenza monitoring are the day the patient is seen, age, sex, diagnosis of influenza-like illness (ILI) or lower respiratory tract infection (LRTI). The reports are usually made on paper forms to form weekly aggregates and sent to a central authority such as the US Centers for Disease Control and Prevention. Paper-based sentinel systems are obviously subject to bureaucratic and postal delay, but we can expect that they will be replaced by real-time, computerized systems. Longer delays are encountered for laboratory-confirmed cases, but recent advances in diagnostic technology have been rapid.

Weekly data make estimates of the growth phase of epidemics appear too short if only the weeks with maximum growth are used, and too long if base-to-peak measures are used. In addition, the amplitude and slope from base to peak are underestimated with weekly data. Aggregation degrades the precision of the temporal "signal" and reduces variability.

Weekly counts for infectious diseases are a legacy of non-electronic recording, mailing, and publishing systems. Models of rapidly spreading epidemics are conceptualised on scales of hours and days: e.g. the periods of latency and infection. It is unsatisfactory to specify models at one time scale and then to estimate the parameters and assess model accuracy using data at a more aggregated scale. Weekly data may also conceal important sequencing changes during the course of an epidemic: e.g. a shift from bronchitis to pneumonia as sequelae of influenza, or a shift from child to adult deaths.

Within the last five years there has been increasing interest in real-time "syndromic surveillance", using indirect indicators. The most publicised is "Google flu trends" which records on a worldwide scale the number of web searches related to influenza and their country of origin. Useful information can also come from telephone triage services, such as enquiries to the British National Health Service Helpline. Many commercial organisations in the health industry such as private hospitals and pharmaceutical retailers have computerised record-keeping, integrated across multiple locations, which may provide information that is more up-to-date than that of the state health providers.

Treating Past Pandemics as Real-Time Problems.

In the past 125 years, there have only been 5 genetic shifts in the human influenza A virus that have led to pandemics. Despite this relatively small number, data from the contemporary registration systems for incidence and deaths have provided important insights into influenza dynamics.

It is surprising, given its importance for epidemic preparedness, that there are relatively few estimates of the Reproductive Number for 1918. Analyses of this pandemic in the United States and North-West Europe produce estimates of around 1.2 to 3.75. These values are surprisingly low when compared with estimates as high as 20 for other influenza epidemics, and when compared with other infectious diseases. It suggests that the 1918 disaster was caused by very high case fatality rates, rather than by extreme transmissibility, and that control might have been feasible with aggressive intervention.

Published reproductive numbers for past epidemics have been estimated *ex post* to maximise knowledge about the process. A complementary *ex ante* approach to the formulation and evaluation of new models designed to exploit real-time data from inter-pandemic periods is to treat previous pandemics and epidemics as if they were happening in real-time. This necessarily requires daily data.

Ex post analyses can use models that are purely statistical, following the time-series tradition, or they can be process models that include specific components for Susceptibility, Exposure, Infection and Recovery (SEIR models). Our *ex post* analysis of daily data for Munich shows that turning points in influenza epidemics can be estimated with confidence intervals as narrow as +/- 1 day. However, for a real-time context it is likely that statistical rather than process models will be preferred, as the latter are usually based on knowledge of the whole epidemic curve.

Data Complications.

In this study we address two complications that are not treated in the established methodology. The first is that we recognise that real-time data may show weekly cycles. It is clear that daily case incidence is depressed at weekends, and on public holidays such as Christmas and New Year - which

occur in the peak influenza season for the Northern Hemisphere. Employee sickness records and pharmaceutical sales may have structural zeroes if the enterprises do not operate on a Sunday. The possibility that morbidity (and sometimes mortality) counts may be depressed at weekends, or that cases could be shifted to Friday or Monday, has been ignored in the literature.

The second problem concerns leads and lags in multiple sources. This issue is explicit in spatial models, but has rarely been addressed for multiple series within one spatial unit. We hypothesise that time-series for employee records and pharmaceutical sales may "lead" doctor visits, which in turn may be lagged by hospital entries and ultimately deaths. The theory of infectious disease modelling states that R_0 is the same for both incidence and mortality if they are correctly recorded. In practice, the use of multiple sources with different reporting characteristics is likely to have contributed to the variability in the estimates of R_0 for a given pandemic.

The Model.

Although we have defined R_0 from a demographic viewpoint, there are a number of basic ways to estimate it, and many variants. Sophisticated deterministic and stochastic models with explicit transmission can be useful for simulations, intervention studies, and scenario building, but they are probably ruled out in a real-time analysis. We assume that daily incidence data are available, but no direct information on laboratory-confirmed cases, recovery, contact or transmission. Additionally, as we are concerned with pandemics we cannot assume that the disease is in an endemic equilibrium.

Therefore in our study we aim to:

1) model a mortality series devoid of the misreporting pattern due to weekends and public holidays;
2) compute an instantaneous R_0, so that additional data could be employed to amend the outcome.

In order to cope with these targets, we assume event counts are indirect observations from a latent distribution, i.e. observed counts are not drawn directly from the distribution of real interest, but rather from another distribution derived from it.

The unknown latent distribution is the continuous series of events over time. A compositional matrix describes how this latent distribution was mixed before generating the data, and it is a characterisation of the mis-registration pattern due to week-ends and public holidays.

The observed counts, therefore, can be viewed as the outcome of a misreporting process that transforms latent series into observed data. For instance, the counts on Friday and Monday may be composed of the actual values on these days plus the misclassified cases from the neighbouring Saturday and Sunday.
Both latent distributions and the elements in the compositional matrix are of interest. Regarding the latent event series, instead of assuming a specific function for its description, we let the data speak by themselves using a smooth curve. Such smooth curves could be considered continuous and this allows an estimation of instantaneous R_0 as the relative derivative of the latent curve with respect to time. The compositional matrix embodies all eventual exchanges between week-days via "misreporting proportions" which could also be estimated.

The composite link model (CLM) of Thompson and Baker (1981) offers an elegant framework to model such a complex structure. Moreover the combination of CLM and penalized likelihood has been already used to estimate smooth latent distributions (Eilers, 2007; Camarda et al., 2008). In this study we employ and modify these methodologies to account for our specific data structure.

Estimating the slope during the rising segment of the epidemic in this way has three advantages. First, one can ignore the decline in the number of susceptibles caused by death and immunity. Second, the slope is independent of the level of under-registration, but only if the level remains constant. Third, the estimate of R_0 and its error interval is not a constant but can vary as data accumulates.

Test Data.

In this study we test our models on two data sets. The first is for the influenza pandemic of 1889-90 as experienced in Munich and Bavaria, an event which is one of the few examples that combine a pandemic, immunological naivete in the population, and multiple sources. We have daily data on reported cases in the city, and series of entries and exits for its two major hospitals. In addition, we have daily sickness records for male employees of a major industry in Bavaria, classified by age, location, and type of job.

The second example is for New York State (excluding the city of New York) in 1918. Daily deaths from influenza and pneumonia separately are given for spatial units, including the cities of Rochester and Buffalo. The additional feature in this case is that distributions of reported times from symptoms to death are available.

It is known that the age-specific response in these two pandemics was different and that the 1889-90 outbreak occurred in an era when influenza had almost disappeared as a cause of morbidity and mortality, so that the proportion of the population with a naïve immune response to influenza was probably much higher than in 1918.

Main reference.

Camarda, C. G., P. H. C. Eilers, and J. Gampe (2008). Modelling General Patterns of Digit Preference. Statistical Modelling 8, 385–401.

Eilers, P. H. C. (2007). Ill-posed Problems with Counts, the Composite Link Model and Penalized Likelihood. Statistical Modelling 7, 239–254.

Thompson, R. and R. J. Baker (1981). Composite Link Functions in Generalized Linear Models. Applied Statistics 30, 125–131.