Predicting the Next Pandemic: An Exercise in Imprecise Hazards

Miķelis Bickis
University of Saskatchewan
bickis@snoopy.usask.ca

Uģis Bickis
Phoenix OHC
bickis@sympatico.ca

Abstract

Influenza pandemics have swept the world numerous times during the last few centuries. Cases of bird flu infecting humans have prompted predictions that we are due for another pandemic soon, but skeptics dismiss such prognostications as panic caused by a misunderstanding of probability. The issue can be reduced mathematically to the question of whether the pandemic process has an increasing, constant, or decreasing hazard function. Historical data on past pandemics can be used to estimate the hazard function using imprecise probabilities, giving upper and lower predictive probabilities of an imminent pandemic, given past waiting times. In order to achieve smoother estimates of the imprecise hazard function, an autocorrelated imprecise Normal prior is proposed.

Keywords. Survival analysis, hazard function, autocorrelated prior.

1 Introduction

Observations of human cases of H5N1 avian influenza in recent years have sparked much discussion in both scientific literature and popular media about the prospects of another flu pandemic. Such pandemics have occurred several times in recent history and concern has been expressed about the prospects of another one. The most devastating occurrence was the Spanish flu of 1918, but lesser pandemics have occured since then, the most recent being an H1N1 strain in 1977 [7]. Experts disagree on the probability of an imminent pandemic, and in an attempt to elicit probabilities, the University of Iowa has even created an online market in avian influenza futures! [1].

One question that arises is whether the probability of an imminent pandemic increases the longer it has been since the last one. On the one hand, it has been argued that pandemics have historically occurred at 20 to 30 year intervals, and given that it has been 30 years since the last one, we are "due" for one. Countering that is the argument [3] that a long waiting time actually makes an imminent pandemic *less* likely since it indicates that the evolutionary course of prospective pathogens has wandered away from genotypes adapted to human transmission. Dismissing both these arguments are individuals with a little learning in probability who proclaim that the probability of a pandemic is unaffected by a long waiting time, since a pandemic is a random event.

A more sophisticated probabilistic view, of course, will acknowledge that any of the three scenarios are logically possible. Suppose that t represents the year of the last pandemic, and let $T+t$ be the year of the next one. If $t+s$ is the current year, then the relevant quantity is the discrete hazard rate

$$\begin{aligned} h(t) &= \Pr\{T+t=t+s+1|T+t>t+s\} \\ &= \frac{\Pr\{T=s+1\}}{\Pr\{T>s\}}, \end{aligned} \tag{1}$$

the conditional probability that it will occur in the next year given that it has not happened yet. Equivalently, one can work with the *instantaneous hazard rate* defined as

$$\lambda(s) = \lim_{\delta\downarrow 0} \frac{\Pr\{T \le s+\delta\}}{\delta \Pr\{T>s\}}, \tag{2}$$

the two concepts being related by

$$S(t) = \Pr\{T>s\} = \exp\left(-\int_0^s \lambda(u)\,du\right) = e^{-\Lambda(s)} \tag{3}$$

and

$$h(s) = 1 - \exp\Big[\Lambda(s) - \Lambda(s+1)\Big] \approx \lambda(s). \tag{4}$$

S is called the survivor function, and Λ is the integrated hazard.

The contrary opinions expressed in the previous paragraph can now be described as believing that the hazard rate is respectively increasing, decreasing, or constant. It is possible to construct probabilistic models that are consistent with any of these viewpoints. While the dynamics of viral evolution is too complex to describe by a simple model, even simplistic models exhibit increasing or decreasing or constant hazards. If the occurrence of a pandemic happens as a result of a number of steps with a strong selective drift, then the hazard will be increasing, since while we are waiting, the virus is getting closer to a pandemic state. On the other hand, if viral evolution is envisaged as a random walk in a space of genotypes then a decreasing hazard would be typical of hitting times in such processes. But if pandemics truly are like a Poisson process, then a constant hazard would be expected.

The purpose of this paper is not to delve into realistic models of viral evolution, nor to propose definitive predictions of an influenza pandemic. Rather, we will examine to what extent one can determine the nature of the hazard function for the pandemic process, based solely on the historical record of past occurrences, and show how principles of imprecise probability cast light on the uncertainty present in such estimates. We will also contrast these methods with classical statistical approaches.

2 Mathematical models and data

According to Patterson [7], influenza pandemics occurred in the following years: 1729, 1732, 1781, 1788, 1830, 1833, 1836, 1889, 1899, 1918, 1957, 1968, and 1977. Some pandemics may have lasted more than a year. We use the first year reported as indicating the beginning of the pandemic.

We consider the pandemics to be a renewal process, in which the time between occurrences are i.i.d. random variables. Thus we are assuming that after each pandemic the virulent strain dies out because of immunity and deaths of hosts, and the evolutionary process to a new strain of pandemic virulence begins anew. We are also assuming no secular trend in the intensity of the process. These assumptions are admittedly simplistic, and may be challenged. Variation of these assumptions would increase the imprecision in the estimates.

Patterson's record gives inter-pandemic periods of 3, 49, 7, 42, 3, 3, 53, 10, 19, 39, 11, and 9 years. To this data we can add the 30 pandemic-free years to the present, which becomes a censored observation.

3 Frequentist analysis

A classical approach to fitting the data is to use the Kaplan-Meier estimator [6]. This allows one to estimate the survivor function S allowing for censoring, but does not directly address the question of increasing or decreasing hazard. We can, however, fit a parametric model

$$\log \lambda(t) = \theta_1 + \theta_2 \log t, \tag{5}$$

which assumes that the interpandemic times have a Weibull distribution. Increasing, constant, and decreasing hazards correspond to positive, zero, and negative values, respectively, of the parameter θ_2.

Estimating the parameters by maximum likelihood gives the estimates $\hat{\theta}_1 = -3.329$, $\hat{\theta}_2 = 0.075$, suggesting a slightly increasing hazard. However, a likelihood ratio test finds that $\hat{\theta}_2$ does not differ significantly from zero, which some people (mistakenly) might interpret as evidence that the hazard is constant. Indeed the 95% confidence set on the parameters establishes only that $-0.44 < \theta_2 < 0.80$, indicating that the data are consistent (under this model) with decreasing, increasing, or a constant hazard. Figure 1 shows the estimated survivor functions for both the Kaplan-Meier and maximum likelihood estimates.

The hazard is more relevant than the survivor function for the predictive probability of an imminent pandemic after a waiting period. As shown in Figure 2, the estimated hazard is nearly constant at about 5%. However, the estimate has considerable uncertainty, which in classical terms is indicated by a confidence set.

Figure 2 also shows a set of hazard functions corresponding to the boundary of the 95% confidence set on the parameters. This envelope well displays the uncertainty, showing both increasing and decreasing hazards that are consistent with the data.

4 Imprecise probability models

While a confidence band on a predictive curve indicates the imprecision, it is not possible to interpret these bounds as predictive probabilities. It is difficult to explain the meaning of the upper envelope in Figure 2 in a way that is both understandable and mathematically correct. Imprecise probability bounds, on the other hand, can honestly be described as upper and lower predictive probabilities.

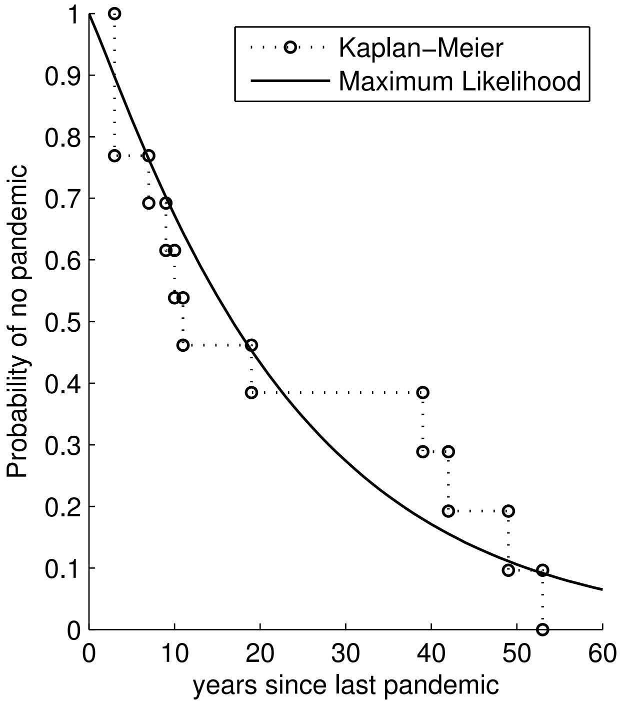

Figure 1: Kaplan-Meier and maximum likelihood estimates of survivor function for pandemic-free periods.

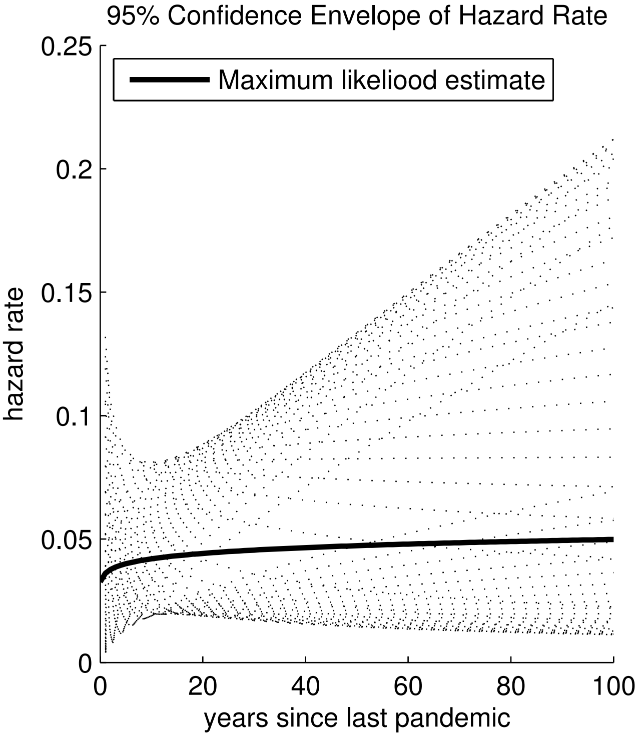

Figure 2: Maximum likelihood estimate of hazard function along with 95% confidence envelope.

4.1 Imprecise Dirichlet and product Beta models

Nonparametric estimates of survivor functions using Walley's [8] imprecise Dirichlet model were discussed by Coolen [4]. However, this model does not give useful descriptions of the hazard function, since upper and lower bounds on the survivor function do not translate into upper and lower bounds on the hazard. One can, however, estimate the hazard function directly.

Suppose that we have k time intervals (which we assume to be equally spaced). Suppose that θ_i represents the conditional probability of a failure (i.e., pandemic) by the end of the ith interval, given that there have been no failures in the preceding $i-1$ intervals. If we have data on m_{i-1} cases in which no failures have occurred, and n_i of them do fail, then this random variable will have a binomial distribution with success probability θ_i. Moreover, since failures in different intervals will be conditionally independent (given survival), the likelihood of a sample will be proportional to

$$\prod_{i=1}^{k}(1-\theta_i)^{m_i}\theta_i^{n_i}. \tag{6}$$

Conjugate to this likelihood would be a product Beta distribution. We can then use, for each interval, an imprecise Beta prior with hyperparameters $\alpha_i\nu$ and $(1-\alpha_i)\nu$ (using the notation of Bernard [2]) where α_i covers the interval $(0,1)$ to give the range of imprecise probabilities. We use the same ν for all intervals, although an argument could be made for varying it. The upper and lower predictive hazards, (i.e., the upper and lower posterior expectations of θ_i) then become $(n_i+\nu)/(n_i+m_i+\nu)$ and $n_i/(n_i+m_i+\nu)$, respectively. The upper and lower survivor functions can then be computed as

$$\overline{\hat{S_i}} = \prod_{j=1}^{i}\left(1-\frac{n_j+\nu}{n_j+m_j+\nu}\right) \tag{7}$$

$$\text{and}\quad \underline{\hat{S_i}} = \prod_{j=1}^{i}\left(1-\frac{n_j}{n_j+m_j+\nu}\right). \tag{8}$$

In the absence of censoring, $m_i = n_i + m_{i+1}$, and as $\nu \to 0$, $\overline{\hat{S_i}}$ becomes the Kaplan-Meier estimator.

Figure 3 shows the upper and lower probabilities of the survival function for both the imprecise Dirichlet model and the product Beta model, as well as the Kaplan-Meier estimator for comparison. Following the suggestion of Walley, we used a value of $\nu = 1$ as the imprecision parameter. It turns out that the upper probabilities of the Dirichlet and product Beta models are identical, whereas the Beta model gives a

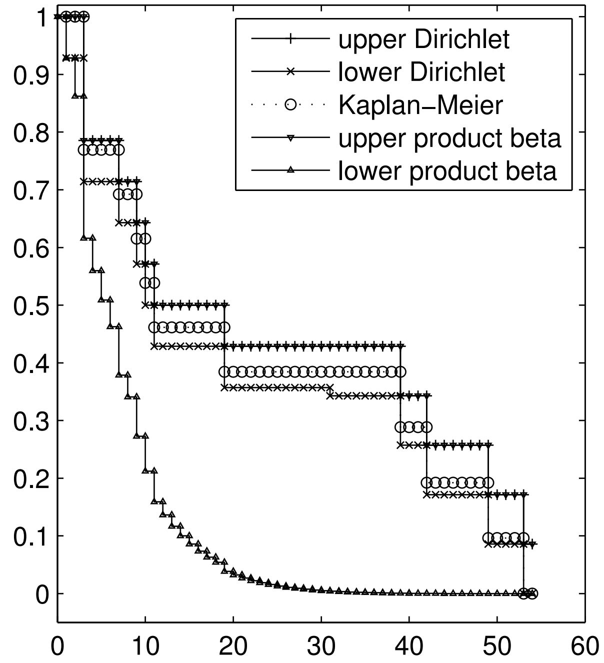

Figure 3: Upper and lower imprecise survivor function, showing both imprecise Dirichlet estimates, and product Betas estimates

substantially smaller lower probability. The Kaplan-Meier estimator lies between the upper and lower probabilities, as was pointed out by Coolen [4].

Figure 4 shows the upper and lower probabilities for the hazard function. Note that when an interval has no occurrences, the lower probability is necessarily zero, while the upper probability can be quite high if the remaining sample numbers are low. The rather jagged shape of the curve can be explained by the fact that if the parameters are independent *a priori* then the form of the likelihood (6) makes them independent *a posteriori* as well.

4.2 Correlated imprecise Normal model

It would be preferable to make use of the prior information that the hazard function would be continuous and fairly smooth. We would not expect drastic changes in the probability of recurrence in a year. Thus, in place of the product Beta model, we are proposing that the prior distribution of the θ's be an autoregressive process. To make this tractable, we use a Gaussian prior on the log-odds.

Specifically, we assume that the $\omega_i = \log\big(\theta_i/(1-\theta_i)\big)$ has *a priori* a Normal distribution with mean μ and variance σ^2. Moreover, we assume that the sequence of ω_i's follow a stationary AR(1) process with autocorrelation ρ.

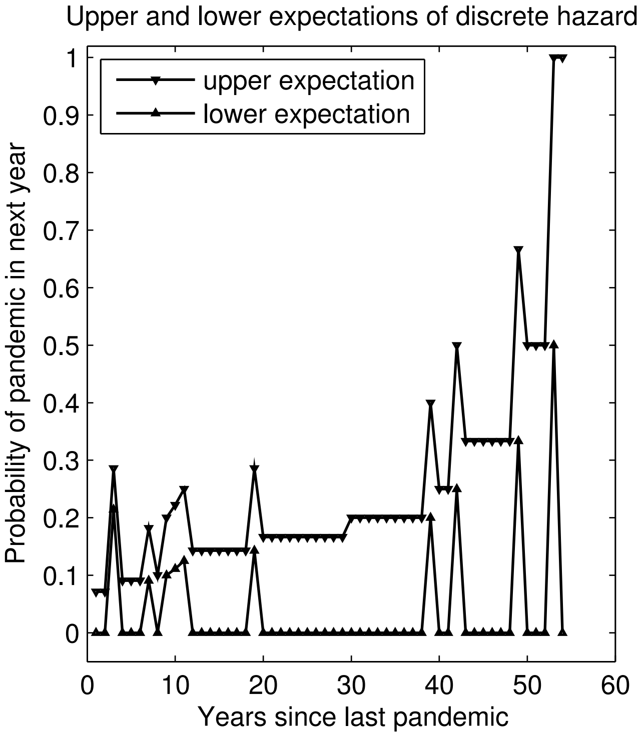

Figure 4: Upper and lower imprecise hazard function.

Using a Beta prior, the distribution of ω_i would be Fisher's-Z [5], which has lighter tails than the Normal. Thus the Normal prior tends to give somewhat less weight to extreme probabilities (which could be viewed as an advantage). Another difficulty is that the posterior distribution is harder to evaluate. Given binomial data of y successes out of n trials, the posterior distribution of ω has density

$$K(\mu,\sigma,n,y)\frac{\exp\left[-\Big(\omega-(\mu+\sigma^2 y)\Big)/(2\sigma^2)\right]}{(1+e^{\omega})^n} \qquad (9)$$

where K is a constant of integration. The posterior mean, (i.e., the predictive probability) appears not to be tractable, but can be computed numerically as

$$K\int_0^1 \exp\left[-\frac{\left(\log\left(\frac{\theta}{1-\theta}\right)-(\mu+\sigma^2 y)\right)^2}{2\sigma^2}\right]\frac{(1-\theta)^{n-1}}{\theta}\,d\theta. \qquad (10)$$

The imprecise Dirichlet model has the property that the prior probabilities are vacuous, but the posterior ones may have some precision. To achieve the same goal with the Normal model requires care. We use the

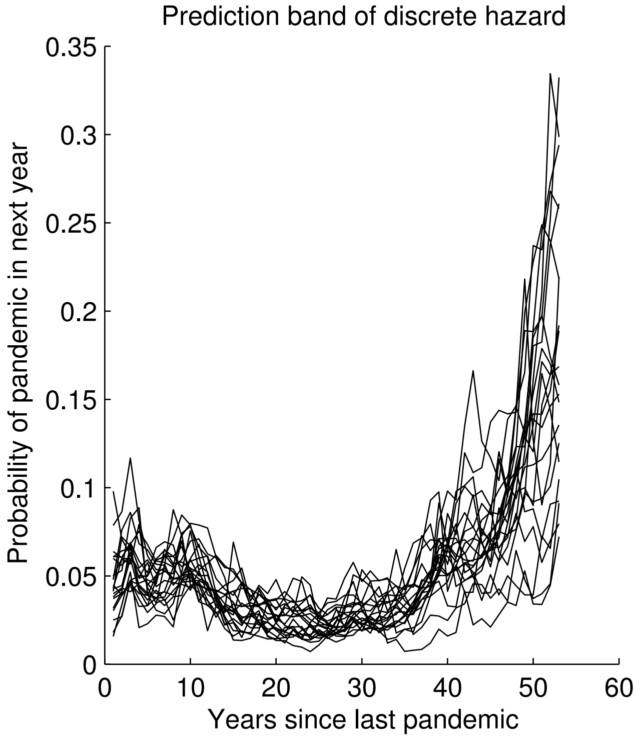

Figure 5: Sampled hazard functions from autocorrelated imprecise posterior

family of Normal distributions where

$$\sigma = \sigma_0 + \tau|\mu|^\gamma \tag{11}$$

where σ_0, τ and γ are viewed as tuning parameters. We use $\sigma_0 = 8/3$ as a value that (with $\mu = 0$) approximates the Beta$(1/2, 1/2)$ density for θ. Thus this symmetric prior distribution represents about the same level of uncertainty as a symmetric Beta distribution with $\nu = 1$. Putting $\tau = \gamma = 0.5$ and letting μ vary from $-\infty$ to ∞ appears to achieve our goal of providing upper and lower probabilities (although more work could be done here).

Extending the integral (9) to the multivariate case seemed intractable, so we estimated the smoothed hazard function using importance sampling. Letting μ vary from -8 to 2, 1000 samples were taken from a Gaussian AR(1) process with mean μ, $\rho = 0.99$ and variance given by (11). For each sample, the likelihood of the observed data was computed. These likelihoods were then used as weights in computing the predictive probabilities of both the hazard and survival functions. The results are shown in Figures 5 and 6. The bundle of curves displays the imprecision in the predictive probabilities.

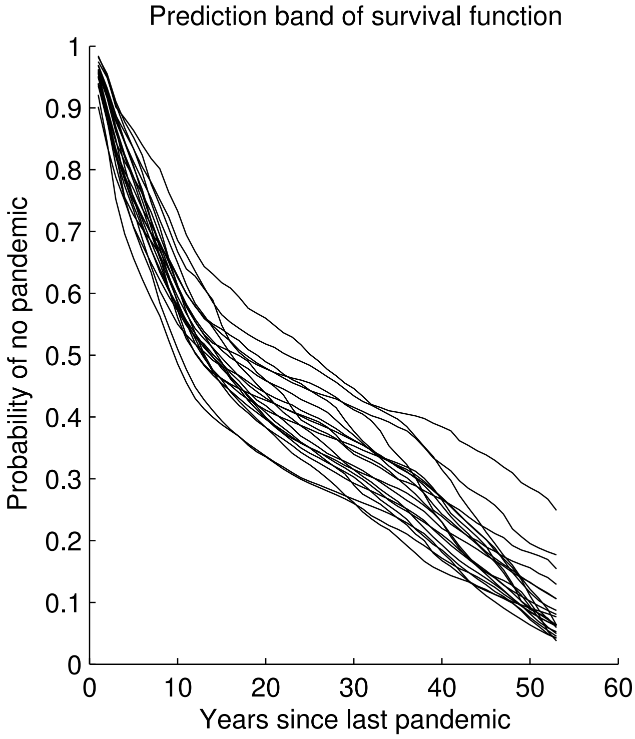

Figure 6: Sampled survival functions from autocorrelated imprecise posterior.

5 Conclusion

From these displays we can see that although a constant hazard can (barely) fit inside this band, there is a rather strong suggestion of an increasing hazard after about 25 years. While this exercise cannot pretend to be the last word on predicting pandemics, it does show how ideas of imprecise probability can focus on realistic understanding of future risks. We hope that imprecise probability methods will be useful in other situations of estimating risks after waiting time. As extension of this work, we intend to examine how the hyperparameters of the stationary Gaussian process affect the performance of the estimates.

References

[1] Anonymous. Avian Influenza – the Iowa Health Prediction Market, web page. `http://flupredictiction.uiowa.edu /fluhome/Market_AvianInfluenza.html`

[2] Jean-Marc Bernard. An Introduction to the Imprecise Dirichlet Model for Multinomial Data. *International Journal of Approximate Reasoning*, 39:123–150, 2005.

[3] Canadian Broadcasting Corporation. Scientific jury still out on prospects of avian flu pandemic, web page. `http://www.cbc.ca/health/story/2006/03/20/avian-flu060320.html`

[4] F. P. A. Coolen. An Imprecise Dirichlet Model for Bayesian Analysis of Failure Data Including Right-Censored Observations. *Reliability Engineering and System Safety*, 56:61–68, 1997.

[5] Harald Cramér. *Mathematical Methods of Statistics.* Princeton University Press, 1946.

[6] E. L. Kaplan and Paul Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[7] K. David Patterson. *Pandemic influenza, 1700-1900 : a study in historical epidemiology* Rowan & Littelfield, 2005.

[8] Peter Walley. Inferences from Multinomial Data: Learning about a Bag of Marbles. *Journal of the Royal Statistical Society*, 58B:3–57, 1996.