Real-time forecasting of infectious disease epidemics

JT Wu *, BJ Cowling

KEY MESSAGES

1. The validity and predictability of three epidemic models were evaluated: a hybrid-type homogeneous stochastic model, an age-structured variant of the previous model, and a power-law logistic model.
2. Reporting rates affect the interpretation of model parameters only but not the performance of parameter estimation or real-time epidemic forecasting.
3. Reliable and precise real-time epidemic forecasting is improbable during the early phase of an epidemic and unlikely to be robust until the epidemic has peaked, when using only epidemic curve data and any of the three models.
4. Robust real-time epidemic forecasting, if possible at all, requires other sources of epidemic data, such as seroprevalence, household transmission data, and phylogenetic data.
5. Epidemiologists and public health policymakers should be aware of these results when using models for real-time epidemic forecasting.

Hong Kong Med J 2018;24(Suppl 6):S26-9
RFCID project number: 12111342

JT Wu, BJ Cowling

School of Public Health, The University of Hong Kong

* Principal applicant and corresponding author: joewu@hku.hk

Introduction

Mathematical modelling of infectious disease has greatly progressed. Much effort has been devoted to statistical inference of model parameters in real time with certain assumptions about the underlying transmission dynamics.[1] In contrast, real-time epidemic forecasting has been understudied. Although robust and well-established forecasting techniques for epidemics remain largely elusive, the ability to predict future incidence is regarded as one of the key functions of epidemic models (especially among non-modellers and policymakers).

Various scientific approaches have been attempted to forecast the course of epidemics of influenza and other directly transmitted diseases. These approaches can be broadly classified into two categories. One relies on a mechanistic model of transmission dynamics, and the other uses statistical extrapolation of epidemic curves. The mechanistic approach explicitly accounts for so-called 'dependent happening', which refers to the dependence of the risk of infection in one individual on the risk in other individuals. These models are built by describing the underlying dynamics of infectious disease transmission to explain the resulting epidemic curves. The other approach includes studies that use a parsimonious but flexible power-law logistic equation to directly fit the flexible parametric model to epidemic curves.[2] Although this approach necessarily disregards dependent happening, mechanistic models can be approximated by a family of logistic equations.

Forecasts of incidence are largely inaccurate and imprecise until the epidemic has peaked and depend strongly on the proportion of population that is susceptible at the beginning of the epidemic. We examined these aspects by comparing three forecasting methods in five case studies to improve the understanding of the feasibility and reliability of these models for real-time epidemic forecast. This study aimed to validate the three forecasting methods by varying precision of data and length of forecast and to assess the predictability of the three forecasting methods by timing and length of forecast.

Methods

We evaluated the following three forecasting methods.

(1) A hybrid-type homogeneous stochastic model.[3] Let R_i be the reproductive number at the beginning of an epidemic, T_g be the mean generation time, S_k be the number of susceptible individuals at the beginning of period k, r_k be the epidemic growth rate in period k, and C_k be the number of cases in period k. Given C_0, ..., C_k, the probability distribution of C_{k+1} is a Poisson distribution with mean $A_k C_k$ where

$$A_k(S_0,R_i,C_k,...,C_0) = \frac{r_k e^{r_k \Delta t}}{r_{k+1}} \frac{e^{r_{k+1}\Delta t}-1}{e^{r_k \Delta t}-1},\ r_k = \ln\left(\frac{S_{k-1}R_i}{S_0}\right)^{\frac{1}{T_g}},\ S_k = S_{k-1}-C_k$$

This model has two parameters: (S_0, R_i).

(2) An age-structured variant of the previous model that is analogous to the model developed by Katriel et al.[4] Let n be the number of age-groups. We assume that given the same level of exposure,

age-group j is a_j times as susceptible to infection as age-group 1 (ie a_1=1). We assume a contact matrix B=[$b_{i,j}$], where $b_{i,j}$ is the contact frequency between age-groups i and j normalised such that the largest eigenvalue of B is 1. We parameterise B using the UK contact matrix from the POLYMOD study.[5] Let $C_{k,j}$ be the number of cases in age-group j on day k. Given $C_{0,j}$, ..., $C_{k,j}$, j = 1, ..., n, the probability distribution of $C_{k+1,j}$ is a Poisson distribution with mean $S_{k,j}R_0a_j\sum_{i=1}^{n}b_{ij}\sum_{l=1}^{k}C_{k-l,i}g_l$, where $S_{k,j} = S_{k-1,j}-C_{k,j}$ is the number of susceptible individuals in age-group j at the beginning of day k, R_0 is the basic reproductive number, and g is the generation time distribution. This model has 2n parameters: (R_0, a_2, ..., a_n, S_1, ..., S_n).

(3) A power-law logistic model[2] in which the cumulative number of cases at time t is $I(t) = \frac{K}{[1+e^{-r(t-t_m)}]^{1/\alpha}}$. The incidence between time t and and $t+\Delta t$ is assumed to be a Poisson distribution with mean $I(t+\Delta t) - I(t)$. This model has four parameters: (K, r, t_m, α).

Parameters are estimated in a Bayesian framework with non-informative flat priors for all parameters. Model validity is assessed using mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE):

$$MAE = \frac{1}{nm}\sum_{j=1}^{n}\sum_{k=1}^{m}\left|E(C_{k,j})-x_{k,j}\right|$$

$$RMSE = \sqrt{\frac{1}{nm}\sum_{j=1}^{n}\sum_{k=1}^{m}\left(E(C_{k,j})-x_{k,j}\right)^2}$$

$$MAPE = \frac{100\%}{nm}\sum_{j=1}^{n}\sum_{k=1}^{m}\left|\frac{E(C_{k,j})}{x_{k,j}}-1\right|$$

where n is the number of age-groups, m is the number of reporting periods over the course of the epidemic, and $E(C_{k,j})$ and $x_{k,j}$ are the number of cases in the jth age-group during the kth period in the model and the observed data, respectively. Model predictability is assessed for short-range and long-range forecasts (Fig 1) by examining (1) the coefficient of variation of forecasted incidence, (2) the percentage of forecasting periods for which the actual incidence lies outside the 95% prediction intervals of the forecasted incidence, and (3) f MAE, defined as the MAE of the model forecast and actual future incidence over the forecast periods.

The first step is to use simulated data to understand the behaviour and performance of the three models before applying them to real epidemic data. We generate the simulated data using a standard age-structured SIR model with a basic reproduction number R_0 of 1.3 and a mean generation time of 3 days (ie epidemiologically similar to a mild influenza pandemic). We assume that the three groups correspond to the 0-19, 20-59, and ≥60 years age-groups in Hong Kong and that group j is j times as susceptible as group 1. To assess the effect of under-reporting on forecast performance, we consider reporting rates of 100% and 5%.

The smallpox dataset contains the monthly number of smallpox cases in the Netherlands from 1870 to 1873 (with no age information). The mean generation time is assumed to be 15 days.

The polio dataset contains the daily number of polio cases in New York City in 1916 (with no age information). The mean generation time is assumed to be 10 days.

The pandemic influenza A/H1N1 dataset contains the daily number of confirmed cases of pandemic influenza A/H1N1 in five age-groups (0-12, 13-19, 20-29, 30-59, and ≥60 years) between 1 September and 15 November 2009 in Hong Kong. This period was selected because schools were closed before 1 September 2009, and the exogenous force of infection from Shenzhen contributed substantially to the transmission of pandemic influenza A/H1N1 in Hong Kong after 15 November 2009 (ie, the transmissibility of the virus was relatively constant during this period.) The mean generation time was assumed to be 3 days.

The SARS dataset contains the daily number of confirmed SARS cases in three age-groups (0-31, 32-49 and ≥50 years) between 15 February and 31 May 2003 in Hong Kong. The age partition was chosen so that the total number of cases in each age-group were similar (which facilitates model fitting). The mean generation time was assumed to be 8 days.

Results

The results of model validation and predictability for the three methods are shown in the Figure. With a reporting rate of 100%, Markov chain Monte Carlo inference did not converge until just before the epidemic peak for models 1 and 2 but converged sooner for model 3, although the model fit was poor before the peak. In terms of model fitting and real-time epidemic forecast, model 2 had the best performance, and model 3 performed better than model 1 until near the end of the epidemic. This was unsurprising, because model 2 is an age-structured variant of model 1 with the correct contact matrix, and model 3 has two more parameters than model 1, giving more flexibility. However, the predictive power of all three models was generally poor; future incidence almost always fell outside the 95% prediction intervals for both short- and long-range forecasts until near the end of the epidemic.

With a 5% reporting rate, the comparative performance of the three methods was similar to that with a 100% reporting rate. Thus, reporting rate had little effect on parameter estimation and epidemic forecasting for all three methods. Further analysis revealed that the mathematical structures of models 1 and 2 were not affected by the incorporation of

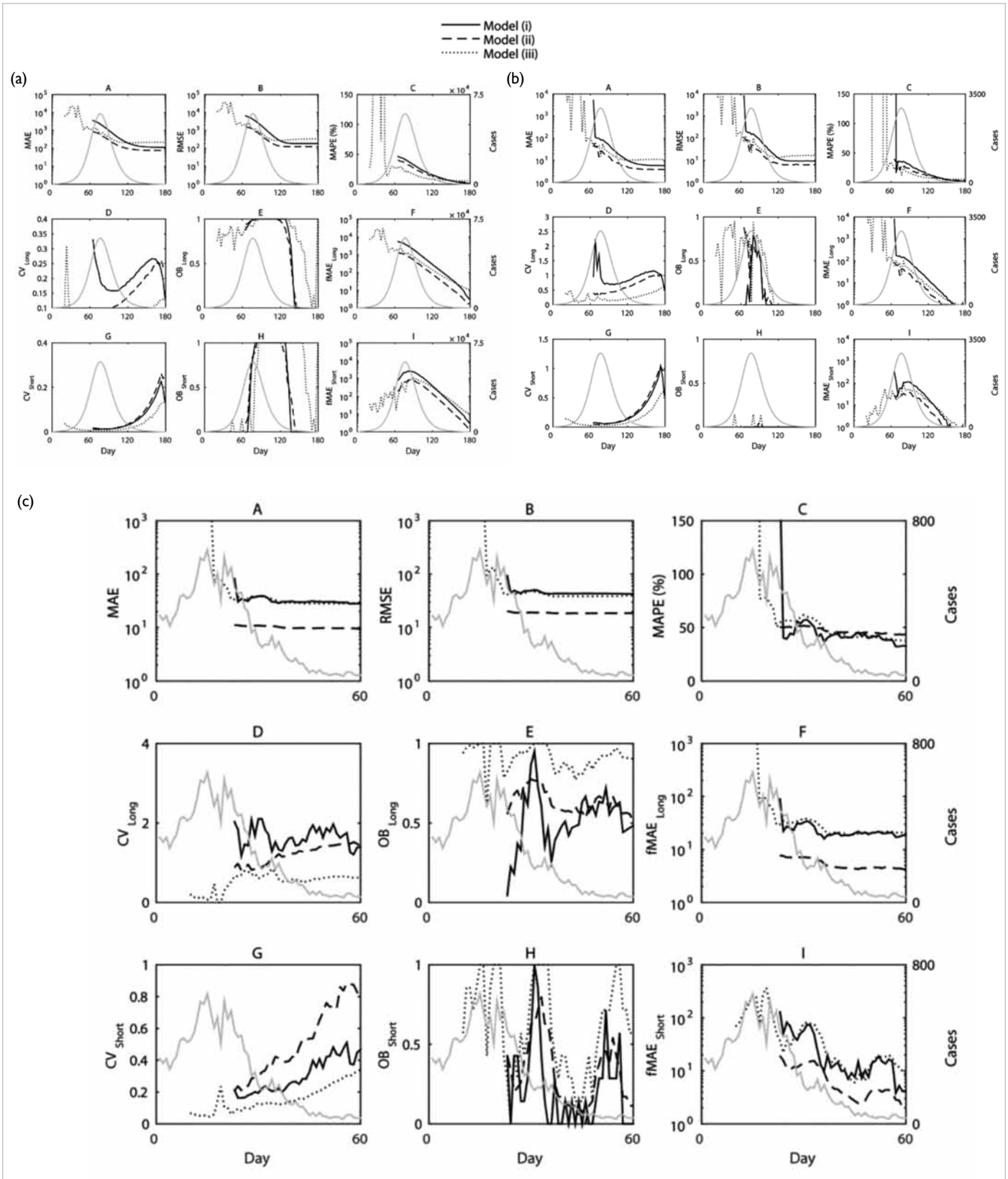

FIG. Real-time epidemic forecasting of the three models in a simulation with (a) a 100% reporting rate, (b) a 5% reporting rate, and (c) the pandemic influenza A/H1N1 case study.

Abbreviations: CV, coefficient of variation; fMAE, forecast mean absolute error; MAE, mean absolute error; MAPE, mean absolute percentage error; OB, out of bounds; RMSE, root mean square error

reporting rate (data not shown).

In the pandemic influenza A/H1N1 case study, models 1 and 2 did not converge until the epidemic peak, but model 3 converged much earlier. Furthermore, the goodness-of-fit of all three models was almost constant over time, indicating that the posterior distributions of the parameters were robust as soon as the epidemic had peaked and the parameters were identifiable. After the peak, model 2 provided the best fit to the data. Model predictability was limited for both long-range and short-range forecasts.

The results in the other three case studies were similar to those of this case study and are documented in the final report of this study.

Discussion

Our results showed that reporting rates affected only the interpretation of model parameters but not the performance of parameter estimation or real-time epidemic forecasting (aside from increased stochasticity because of lower case counts). In all five case studies, the parameter values were largely not identifiable (ie Markov chain Monte Carlo did not converge) for models 1 and 2 until or even after the epidemic peak.

Reliable and precise real-time epidemic forecasting is improbable during the early phase of the epidemic and unlikely to be robust until the epidemic has peaked when using only epidemic curve data and any of the three models. Robust real-time epidemic forecasting, if possible at all, requires other sources of epidemic data, such as seroprevalence, household transmission, and phylogenetic data. Model predictability should be evaluated not only by computing simple error measures between actual and forecasted incidence but also by interpreting these measures in the context of the forecast's level of uncertainty (the wider the prediction interval, the less useful the forecast, but the more likely that the actual incidence falls within the prediction interval). Epidemiologists and public health policymakers should be aware of these drawbacks when using models for real-time epidemic forecasting.

Acknowledgement

This study was supported by the Research Fund for the Control of Infectious Diseases, Food and Health Bureau, Hong Kong SAR Government (#12111342).

References

1. Grassly NC, Fraser C. Mathematical models of infectious disease transmission. Nat Rev Microbiol 2008;6:477-87.
2. Hsieh YH, Cheng YS. Real-time forecast of multiphase outbreak. Emerg Infect Dis 2006;12:122-7.
3. Nishiura H, Chowell G, Heesterbeek H, Wallinga J. The ideal reporting interval for an epidemic to objectively interpret the epidemiological time course. J R Soc Interface 2010;7:297-307.
4. Katriel G, Yaari R, Huppert A, Roll U, Stone L. Modelling the initial phase of an epidemic using incidence and infection network data: 2009 H1N1 pandemic in Israel as a case study. J R Soc Interface 2011;8:856-67.
5. Mossong J, Hens N, Jit M, Beutels P, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 2008;5:e74.