Epidemics 17 (2016) 19–26

Contents lists available at ScienceDirect

Epidemics

journal homepage: www.elsevier.com/locate/epidemics

Data-driven outbreak forecasting with a simple nonlinear growth model

Joceline Lega*, Heidi E. Brown

University of Arizona, Tucson, AZ, USA

ARTICLE INFO

Article history:
Received 8 July 2016
Received in revised form
20 September 2016
Accepted 9 October 2016
Available online 11 October 2016

Keywords:
Infectious disease outbreaks
Mathematical model
Surge capacity
Chikungunya virus infection

ABSTRACT

Recent events have thrown the spotlight on infectious disease outbreak response. We developed a data-driven method, *EpiGro*, which can be applied to cumulative case reports to estimate the order of magnitude of the duration, peak and ultimate size of an ongoing outbreak. It is based on a surprisingly simple mathematical property of many epidemiological data sets, does not require knowledge or estimation of disease transmission parameters, is robust to noise and to small data sets, and runs quickly due to its mathematical simplicity. Using data from historic and ongoing epidemics, we present the model. We also provide modeling considerations that justify this approach and discuss its limitations. In the absence of other information or in conjunction with other models, *EpiGro* may be useful to public health responders.

1. Introduction

As infectious diseases are identified for the first time or emerge in new populations, researchers increasingly use mathematical models to describe observed patterns and to plan and evaluate public health responses (Anderson and May, 1992; Grassly and Fraser, 2008; Keeling and Danon, 2009; Anderson et al., 2015). These models vary in complexity and scale, from simple compartmental models (Hethcote, 2000) to complex stochastic agent-based and metapopulation approaches that include external information like transportation networks (Rvachev and Longini, 1985; Hufnagel et al., 2004; Eubank et al., 2004; Ferguson et al., 2006; Balcan et al., 2010; Ajelli et al., 2010; Van den Broeck et al., 2011). The latter have been shown to efficiently capture the real-time spread of epidemics (Tizzoni et al., 2012), but often require large amounts of information. Key parameters need to be estimated from epidemiological data, which may be accomplished by maximum likelihood estimation (Ionides et al., 2006; Bretó et al., 2009; King et al., 2015) or data assimilation (Rhodes and Hollingsworth, 2009; Shaman and Karspeck, 2012). However, for newly emerging infections or when estimating the impact of bioterrorism events (Walden and Kaplan, 2004; Rotz and Hughes, 2004), such information may not always be available. Sometimes, the community is able to quickly compile and share epidemiological parameters, as was for instance the case for the devastating 2014/2015 Ebola outbreak (Van Kerkhove et al., 2015; Chowell et al., 2014). It is nevertheless expected that model choices reflect the balance between data availability and the needs of the public health community (Keeling and Danon, 2009). Moreover, since the accuracy of predictions depends heavily on modeling assumptions (Keeling and Danon, 2009; Wearing et al., 2005), it is also important to balance the need for detailed, realistic models against limitations in parameter information (May, 2004).

Knowing how many cases to expect, as well as when they will peak, before an outbreak has run its course is central to preparing a public health response (Flu Activity Forecasting Website Launched, 2016). Entire epidemiological curves can often be fitted with standard functions, such as for instance a logistic curve or the Richards model (Tjørve and Tjørve, 2010; Peleg and Corradini, 2011; Wang et al., 2012; Ma et al., 2014), but are only effective late into the outbreak. Conversely, time series approaches allow forecasting, but are considered accurate only for short-term prediction. For instance, using only case data and an autoregressive integrated moving average (ARIMA) model, researchers were able to forecast hospital bed utilization during the severe acute respiratory syndrome (SARS) outbreak in Singapore up to three days forward (Earnest et al., 2005). Additional information is usually required for longer forecasts (see e.g. 3-month dengue forecasting using climate data (Gharbi et al., 2011)), limiting the utility of such approaches for newly emerging diseases, when many associated risk factors are still unknown.

* Corresponding author at: Department of Mathematics, University of Arizona, 617 N. Santa Rita Avenue, Tucson, AZ 85721, USA.
E-mail address: lega@math.arizona.edu (J. Lega).

http://dx.doi.org/10.1016/j.epidem.2016.10.002

We identify a simple property common to the epidemiological curves of many outbreaks and explore the modeling implications of this finding. In particular, it allows us to describe the course of each outbreak in terms of a very simple model, whose two parameters can be extracted from epidemiological data. This is different from estimating disease transmission rates since, for instance, knowledge of the model discussed in this article is not sufficient to recover the parameters (e.g. R_0) of a simulated epidemic that follows the SIR (Susceptible – Infected – Removed) dynamics. We present an automated parameter extraction method that allows us to explore the applicability of the method to a variety of different outbreaks and, more importantly, explain how the model may be used to forecast the scope of ongoing outbreaks, including those of some vector-borne diseases.

2. Methods

Our general methodology is described in Fig. 1. Starting from reported epidemiological data, we consider the *cumulative number of cases*, C, and numerically produce a smooth interpolation of its evolution (panel 1). Data collection procedures for the examples discussed in this article are given in Technical Appendix 1 in Supplementary Material. We then use this smoothed data to estimate incidence, G, as described in Technical Appendix 2 (see Supplementary Material). The crucial point of our approach is that rather than plotting C as a function of time, we plot the estimated incidence G, as a function of cumulative cases, C, G(C) (panel 2). For many outbreaks, the graph of G as a function of C has a single "hump" and can, at first order, be approximated by an inverted parabola (panel 3). This inverted parabola, whose equation contains two parameters, defines a simple model for the evolution of the outbreak, which can be used to predict future number of cases given an initial condition (panel 4). We developed a method, detailed in Technical Appendix 3 in Supplementary Material, that automatically associates a parabola to available epidemiological data of one-wave outbreaks. It works on partial (for ongoing outbreaks) or full (for outbreaks that have completed their course) data sets and proceeds as follows: rather than attempting to estimate the parabola parameters from the cumulative epidemiological curve, we fit the graph of G(C) to its parabolic approximation and the graph of C(t) to its corresponding time course, *simultaneously*. Doing so therefore demands that the two unknown parameters describing the parabola be chosen to provide good approximations of two different (albeit related) plots. This approach is easily applicable to ongoing outbreaks for which limited data are available, and can therefore be used for forecasting.

3. Results

3.1. Robustness over multiple systems

The proposed approach applies to one-wave outbreaks of multiple diseases and sizes, as illustrated in Fig. 2 and supported by our analysis of a variety of epidemiological curves (see additional Appendix figures). The model was tested in detail on nine one-wave outbreaks: 2014–15 chikungunya outbreaks in the Dominican Republic (Fig. 2A), Guadeloupe (Fig. A1), and Dominica (Fig. A2); 2014–15 Ebola outbreaks in Guinea (Fig. A3), Liberia (Fig. A4), and Sierra Leone (Fig. A5); 2008 outbreak of Salmonella SaintPaul in the US (Fig. A6); 2008 outbreak of gastroenteritis in Majorca (Fig. 2B); and 2009 outbreak of H1N1 in Canada (Fig. A7), as well as on one two-wave outbreak of pertussis (2011–12 in the state of Washington, US; Fig. 4). The parabolas plotted in the figures were selected using the automated parameter approximation method. Inspection of these plots reveals that they capture the time course of the cumulative number of cases fairly well (right panel of each figure and as depicted in panel 5 of the schematic of Fig. 1). For very noisy data (e.g. left panels of Figs. A3–5 for Ebola), the chosen parabola nicely interpolates through widely oscillating reported incidence data. The peak incidence (maximum of the blue solid curve on the left panel of each figure) is typically higher than the maximum M of each parabola (Figs. 2B, 4, A1–A5) and may not occur at the same value of the cumulative number of cases. The time frame for the peak of the outbreak (that is when the cumulative curve shown on the right panel of each figure is the steepest), as well as the duration of the entire outbreak (when incidence returns to values close to zero) are however reasonably well captured.

For these reasons, we expect the parabolic model to describe general trends of one-wave outbreaks, such as order-of-magnitude estimates for their final number of cases, duration, and time frame of peak incidence. These statements are made more quantitative below.

A reason for the versatility of this approach is that the parabolic approximation is also "hidden" in the standard SIR model. Fig. 3 presents simulations of this model for small and large values of $R_0 > 1$. The left panel of each row shows the time course of S, I, and R scaled to the total population $N = S + I + R$, and the right panel shows a numerical evaluation of the scaled incidence G/N as a function of scaled cumulative cases C/N (solid curve), together with two parabolic approximations P_1 and P_2. In the context of the SIR model, $C = R + I$ is the total number of cases and its rate of change $G = dC/dt$ is incidence. It is clear from these simulations that for both values of R_0, the SIR model displays the one-hump behavior seen in the graphs of G(C) obtained from outbreak data, and the graph of G as a function of C is very close to an inverted parabola.

The two parabolic approximations P_1 and P_2 plotted in Fig. 3 cross the horizontal axis at $C = 0$ and $C = C_0$, where C_0 is the final number of cases in the model and can be numerically estimated by solving a transcendental equation (details are in Technical Appendix 4 in Supplementary Material). The maximum M_1 of parabola P_1 is a numerical evaluation of the maximum of G. The maximum M_2 of parabola P_2 is equal to $\gamma\, C_0\, (R_0\text{-}1)/4$, based on a theoretical justification also provided in Technical Appendix 4 in Supplementary Material. Both parabolic approximations are very good, but P_1, which estimates M from the data, gives a better fit than P_2. The method proposed in this article proceeds in a similar way: it numerically estimates values of C_0 and M that provide as good a parabolic fit of G as possible, given the available data.

The approach can be extended to multiple-wave outbreaks, in which case a parabola, or a piece thereof, is fit to each wave. The number of waves in any outbreak may easily be identified by plotting G as a function of C, where C, the cumulative number of reported cases, is known as a function of time. Fig. 4 provides an example of a two-wave outbreak, with incomplete data. In this case, the growth rate G is modeled as a piecewise parabolic function of C. Our approach predicts a total of 4232 cases by the end of the outbreak. An article published in 2014 (Bowden et al., 2014) mentions over 4900 cases reported by the end of 2012.

3.2. Robustness over noisy data sets

The fit of G to a parabola, or to pieces thereof, does not have to be perfect to produce good, order of magnitude estimates of the time evolution of the cumulative number of cases of an outbreak. In particular, fluctuations in G (incidence) do not change the overall one- or two-hump behavior of the curve. To make this statement more quantitative, we assessed whether adding a small amount of noise (see Technical Appendix 5 in Supplementary Material for noise generation) to the data significantly affected the outcome of the automated parameter estimation procedure. Specifically, we found that the standard deviation of the distribution of estimates