Mathematical Approaches for Simulating Epidemic Progression: Addressing Limitations of the Linear Chain Trick in ODE Models

Ningrui Xie
Supervisor: Dr. Jonathan Dushoff

Abstract

Mathematical models play a pivotal role in epidemiology by offering insights into the dynamics of infectious disease spread and enabling predictions of future trends. Yet the commonly used SIR model relies on the unrealistic assumption of exponentially distributed infectious stage durations. One strategy used by modelers to solve this problem is to subdivide the infectious stage into an integer number of substages, each having exponentially distributed substage durations with the same recovery rate (SI^nR model). This technique, often referred to as the Linear Chain Trick (LCT), enables the modeling of the entire infectious period durations as Erlang-distributed, aligning more closely with biological realism. However, the LCT approach has limitations, including cumbersome shape parameter determination process and discrete parameter values. In this study, we propose a novel model that assumes a geometrically distributed recovery rate for each infectious substage. We aim to assess its performance, aligning with the established SI^nR model, determining optimal parameters, and evaluating its fitness with real-world data. This approach holds the potential to enhance model flexibility and to develop user-friendly tools for effective epidemic modeling and prediction.

Introduction

In the realm of epidemiology, mathematical models serve as a powerful tool for understanding and effectively managing the spread of infectious disease. They enable us to identify critical factors influencing transmission and predict future trends.

Among the prominent models employed in epidemiology studies, the SIR compartmental model stands as a prevalent choice [1][2]. This model categorizes the population into three compartments representing Susceptible, Infectious, and Recovered individuals. Central to this model is the assumption that the rate at which individuals exit the infectious compartment follows an exponential distribution with a mean time of $\frac{1}{\gamma}$, where γ represent the individual recovery rate [3]. However, this assumption, although mathematically convenient, deviates from the biological reality, as it fails to account for the duration of time individuals already spend in the infectious stage. A more biologically plausible approach would involve a probability distribution where the likelihood of exiting the infectious stage begins small but gradually increases as time approximates the mean duration of infection $\frac{1}{\gamma}$, eventually approaching zero as time extends infinitely. In other words, the probability distribution of the duration required to exit the infectious stage should exhibit a pronounced central tendency [4][5][6][7].

To address this biological incongruity and to offer a more practical way of constructing ODE models, the Erlang distribution [8][9][10], often described as the gamma distribution with integer shape parameters, becomes an instrumental choice for defining the duration of the infectious stage [11][12]. **[JD:** *Why do you say "often described"? It's just a definition. Try to keep your language clean and clear.*] This distribution allows us to employ the concept of subdividing the single infectious stage into n substages, each following an exponential distribution with a recovery rate of $n\gamma$ ($\gamma_i = n\gamma$). The subdivision is made feasible by the property of the Erlang distribution, which is equivalent to a sequence of independent and identically distributed exponential distributions [13]. **[JD:** *What is meant by "made feasible by"? We could literally do it anyway. It "leads to an Erlang", I guess would be better language.*] We represent the Erlang-distributed SIR model as SI^nR. The probability density of Erlang distribution with shape parameter n and scale parameter $n\gamma$ have the following form:

$$f(x; n, n\gamma) = \frac{(n\gamma)^n}{(n-1)!} x^{n-1} e^{-n\gamma x} \quad x > 0, n \in \mathbb{N}$$

Note that using $n\gamma$ as the substage recovery rate preserves the mean recovery period ($\frac{1}{\gamma}$):

$$\frac{1}{\gamma} = \sum_{i=1}^{n} \frac{1}{\gamma_i} = \sum_{i=1}^{n} \frac{1}{n\gamma}$$

This approach, often referred to as the Linear Chain Trick (LCT)[14], is a powerful technique used to transform continuous-time stochastic state transition models, where an individual's time spent in a specific state follows an Erlang distribution, into mean-field ODE models [15]. Importantly, predictions obtained from the SI^{n}R model can exhibit significant divergence from those of the original SIR model, even when the mean infectous stage duration remains the same. This disparity serves to rectify the tendency of overoptimistic predictions that suggest minimal control measures are sufficient to mitigate an epidemic in the SIR model[16].
[JD: *Try to use simpler language – "disparity serves to rectify"?*]

Nonetheless, the utilization of the Linear Chain Trick to generate Erlang-distributed stage durations comes with certain limitations. Firstly, the process of determining the appropriate number of substages (shape parameter, n) is cumbersome and time-intensive. Previous research employing this technique necessitated data fitting for each model with varying substage counts [16][17]. Prior studies have explored methods for estimating the number of states in the linear chain through techniques like profile likelihood [18] and model reduction [19]. However, the inherent rigidity in the model's parameters remains a substantial limitation. As the infectious stage in the SI^{n}R model must be divided into an integer number of substages for incorporation into the ordinary differential equations (ODE), the parameters of interest, specifically the recovery rate of each substage ($n\gamma$) and the number of substages (n), are inherently governed by discrete values. These constraints substantially restrict the model's adaptability and performance in practical applications.

In this study, we aim to address these limitations by introducing a geometrically distributed recovery rate ($\gamma_i = ar^{i-1}$) from each infection substage rather than a constant rate of $n\gamma$. This shift in perspective eliminates the need for using integers as the shape parameter to control the distribution, allowing us to maintain a constant value for n while leveraging a

and r to modify its shape, thereby transitioning our parameters of interest from discrete n and $n\gamma$ to continuous a and r. We refer this novel model as the SIgR model. In the initial phase of our study, we will test the performance of the proposed model by assessing its ability to align with the established SI^{n}R model. This will involve devising a method to choose values of a and r to accurately match with the mean (M) and quadratic coefficient of variation (κ) of the SI^{n}R model across various substages, and evaluating the shape of the probability distribution for the duration of the infectious stage. We will subsequently conduct an empirical validation of our model through the process of data fitting. This evaluation will encompass aspects such as the model's alignment with real-world data and its predictive capabilities in projecting future trends. We will carry out this assessment alongside the established SI^{n}R model using the identical dataset to analyze the effectiveness of the SIgR model. We anticipate that this approach will enhance the model's flexibility and utility, thus paving the way for the development of user-friendly software tools that empower modelers to efficiently select and employ the most suitable models for describing and predicting future epidemic scenarios.

Research Design

We begin our study by examing the distribution of the stage durations [**JD:** *What does this mean? I think we're interested in the overall duration, not the stage durations.*] within an SI^{n}R model to determine whether it conforms to an Erlang distribution when dividing the infectious stage into a sequence of n substages, each exponentially distributed with a mean recovery rate of $n\gamma$. [**JD:** *This is a good step for us to confirm what we think is already known, right? We're not determining that it conforms, just that we're doing it right, I think.*] To achieve this, a simplified I^{n}R model has been created, assuming a fixed number of infected individuals at time 0, with no new infections occurring throughout the study period, and

confers permanent immunity upon recovery:

$$\begin{aligned}
\frac{dI_1}{dt} &= -n\gamma I_1 \\
\frac{dI_2}{dt} &= n\gamma I_1 - n\gamma I_2 \\
&\vdots \\
\frac{dI_n}{dt} &= n\gamma I_{n-1} - n\gamma I_n \\
\frac{dR}{dt} &= n\gamma I_n
\end{aligned}$$

We employed the R package 'deSolve' to solve the I^nR model, enabling us to examine how the number of infected and recovered individuals changes over time. By setting the initial number of infected individuals to 1, we derived the cumulative distribution function (CDF) for both infected and recovered individuals. [**JD:** *I think we have the CDF for recovered individuals only. If I were being more technical, what we're really interested in is the CDF of who has* left *the infectious class. It corresponds to recovered in this case, but it would be good to consider the question more generally.*] Taking the derivative of the CDF functions obtained form the model would give us the probability density function (PDF), which facilitates a comparison with the Erlang distribution, characterized by shape parameter (number of substages) n and scale parameter (recovery rate of each substage) $n\gamma$. The outcomes of this analysis will serve to empirically validate the notion that the sum of an integer number of independent exponentially distributed random variables follows an Erlang distribution [13]. After achieving concordance between the Erlang distribution and the infectious stage duration distribution defined by the I^nR model, we will use these two interchangeably.

Established epidemic models featuring Erlang-distributed infection stage durations use a constant recovery rate $n\gamma$ within each substage. In our next phase, we intend to replace this constant rate with a geometrically distributed sequence $\gamma_i = ar^{i-1}$. We begin by assessing if the modified model retains the conventional Erlang-distributed infectious stage duration shape of the while preserving M and κ. For models involving geometric infection substage durations (SIgR), the parameters of interest shift from n and $n\gamma$ to a and r.

Our initial approach involves fixing the number of substages for the geometric distribution (n_{PE}) at 12 and determining the optimal r value to approximate the shape of the Erlang distribution with various infection stages (n_E). However, the presence of the power n introduces complexity when attempting to explicitly determine r and a for achieving concordance in M and κ between the infectious stage defined by SIgR and the Erlang distribution, shown as follow:

$$M: \quad \underbrace{\frac{1}{\gamma}}_{\mathrm{SI^nR}} = \underbrace{\frac{\frac{1}{a}(\frac{1}{r^{n_{PE}}}-1)}{\frac{1}{r}-1}}_{\mathrm{SIgR}}$$

$$\kappa: \quad \underbrace{\frac{1}{n_E}}_{\mathrm{SI^nR}} = \underbrace{\frac{\frac{\frac{1}{a^2}(\frac{1}{r^{2n_{PE}}}-1)}{\frac{1}{r^2}-1}}{M^2}}_{\mathrm{SIgR}}$$

Therefore, we plan to address M and κ separately. The approach involves fixing M and determining the range of r that satisfies the equalization of κ. To maintain the average recovery time as $\frac{1}{\gamma}$, certain constraints must be met:

$$a = \frac{\gamma(\frac{1}{r^{n_{PE}}}-1)}{\frac{1}{r}-1}$$

Determining the range of values for the variable r will faciliate the process of identifying the specific r value, within that range, at which K equals $\frac{1}{n_E}$. [**JD:** *It would be great to rewrite the math here to be simple and clear, matching what we did when we were coding together. The key points are: there are simple equations; we can match κ without worrying about the mean; we can the scale the mean to the desired value without changing κ. Next of course we want robust functions that can validate that the simulation gives us the desired values of μ and κ, and then we want to compare what the pseudo-Erlang and Erlang distributions look like.*]

Upon successfully identifying a and r values that match M and κ of the infection stage durations distribution between the two models, we will take our approach further. Given a predetermined number of substages, a desired κ value, and mean, we aim to numerically

compute the r value that matches κ. This approach will result in a novel methodology for characterizing the stage distribution, which offers increased flexibility compared to the conventional approach using a integer number of substage with identical rate, as the parameters a and r in the new model will be continuous rather than discrete, providing more refined control over the shape of the distribution.

Following the initial theoretical evaluation of the viability of the SIgR model, our research will progress to the empirical phase, where data fitting will be undertaken to evaluate the model's efficacy. The dataset under consideration encompass disease outbreak data, vaccination records, and case-specific information. The fitting process entails the adjustment of model parameters to achieve the best fit. The assessment of model fitness will be executed through methodologies such as maximum likelihood estimation and Bayesian inference. Subsequently, the goodness-of-fit of each model will be calculate to compare their respective performance and suitability for the given dataset. [**JD:** *I am not sure why you are talking about goodness-of-fit here; it's OK to be very brief, more important to be clear than to write long sections.*]

Concluding Remarks

Our proposed projects and experiments will assess the effectiveness of our model. Eventually, these methods may making it easier for modelers to fit model with flexible time distributions in an ODE framework. This research holds the potential to guide more effective public health responses and decision-making during future epidemics.

References

[1] William Ogilvy Kermack and A. G. McKendrick. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, 115(772):700–721, 1927.

[2] R. M. Anderson, R. M. May, M. C. Boily, G. P. Garnett, J. T. Rowley, and R. M. May. The spread of hiv-1 in africa: sexual contact patterns and the predicted demographic impact of aids. *Nature*, 352:581–589, 1991.

[3] Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, 2000.

[4] Philip E. Sartwell. The distribution of incubation periods of infectious disease. *American Journal of Epidemiology*, 51(3):310–318, 1950.

[5] R.E. Hope Simpson. Infectiousness of communicable diseases in the household. *The Lancet*, 260(6734):549–554, 1952.

[6] Norman T. J. Bailey. A statistical method of estimating the periods of incubation and infection of an infectious disease. *Nature*, 174:139–140, 1954.

[7] K. J. Gough. The estimation of latent and infectious periods. *Biometrika*, 64(3):559–565, 1977.

[8] Dorothy Anderson and Ray Watson. On the spread of a disease with gamma distributed latent and infectious periods. *Biometrikae*, 67(1):191–198, 1980.

[9] A. L. Lloyd. Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods. *Biometrikae*, 268(1470), 2001.

[10] Junling Ma and David J. D. Earn. Generality of the final size formula for an epidemic of a newly invading infectious disease. *Bull. Math. Biol*, 68:679–702, 2006.

[11] Olga Krylova and David J. D. Earn. Effects of the infectious period distribution on predicted transitions in childhood disease dynamics. *Journal of The Royal Society Interface*, 10(84), 2013.

[12] Hanh T.H Nguyen and Pejman Rohani. Noise, nonlinearity and seasonality: the epidemics of whooping cough revisited. *J. R. Soc. Interface*, 5(21):403–413, 2007.

[13] Charles Therrien and Murali Tummala. *Probability for Electrical and Computer Engineers*. CRC Press; 1st edition, 2011.

[14] Norman MacDonald. *Time Lags in Biological Model*. Springer Science, 1978.

[15] Paul J. Hurtado and Adam S. Kirosingh. Generalizations of the 'linear chain trick': incorporating more flexible dwell time distributions into mean field ode models. *J. Math. Biol*, 79:1831–1883, 2019.

[16] Helen J Wearing, Pejman Rohani, and Matt J Keeling. Appropriate models for the management of infectious diseases. *PLoS Med*, 2(7):e174, 2005.

[17] Paul J. Hurtado and Cameron Richards. Building mean field ode models using the generalized linear chain trick and markov chain theory. *Journal of The Royal Society Interface*, 15(sup1):S248–S272, 2021.

[18] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.

[19] Tim Maiwald, Helge Hass, Bernhard Steiert, Joep Vanlier, Raphael Engesser, Andreas Raue, Friederike Kipkeew, Hans H. Bock, Daniel Kaschek, Clemens Kreutz, and Jens Timme. Driving the model to its limit: Profile likelihood based model reduction. *PLoS ONE*, 11(9):e0162366, 2016.