Physica A 494 (2018) 288–293

Contents lists available at ScienceDirect

Physica A

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

Modeling influenza-like illnesses through composite compartmental models

Nir Levy, Michael Iv, Elad Yom-Tov *

Microsoft Israel, Herzeliya 46733, Israel

HIGHLIGHTS

- High dimensional extension to the SIR model of infectious disease to is proposed.
- Aggregate data of multiple pathogens is fit to the model using matching pursuit.
- Applicability of the method to seasonal influenza-like illnesses is demonstrated.
- Components derived from the model are matched to known viral profiles.

ARTICLE INFO

Article history:
Received 25 June 2017
Received in revised form 6 December 2017
Available online 11 December 2017

Keywords:
Influenza-like illness
SIR models
Social media

ABSTRACT

Epidemiological models for the spread of pathogens in a population are usually only able to describe a single pathogen. This makes their application unrealistic in cases where multiple pathogens with similar symptoms are spreading concurrently within the same population. Here we describe a method which makes possible the application of multiple single-strain models under minimal conditions. As such, our method provides a bridge between theoretical models of epidemiology and data-driven approaches for modeling of influenza and other similar viruses.

Our model extends the Susceptible–Infected–Recovered model to higher dimensions, allowing the modeling of a population infected by multiple viruses. We further provide a method, based on an overcomplete dictionary of feasible realizations of SIR solutions, to blindly partition the time series representing the number of infected people in a population into individual components, each representing the effect of a single pathogen.

We demonstrate the applicability of our proposed method on five years of seasonal influenza-like illness (ILI) rates, estimated from Twitter data. We demonstrate that our method describes, on average, 44% of the variance in the ILI time series. The individual infectious components derived from our model are matched to known viral profiles in the populations, which we demonstrate matches that of independently collected epidemiological data. We further show that the basic reproductive numbers ($R0$) of the matched components are in range known for these pathogens.

Our results suggest that the proposed method can be applied to other pathogens and geographies, providing a simple method for estimating the parameters of epidemics in a population.

* Corresponding author.
E-mail addresses: nirlevy@microsoft.com (N. Levy), t-miiv@microsoft.com (M. Iv), eladyt@microsoft.com (E. Yom-Tov).

https://doi.org/10.1016/j.physa.2017.12.052
0378-4371/

1. Introduction

Compartmental models, first suggested by Hamer [1] and later developed by Kermack and McKendrick [2] describe the spread of an infection through the interaction between the parameters of a pathogen and three populations: susceptible individuals, infective individuals and recovered individuals. These models are commonly known as susceptible–infected–recovered (SIR) models. SIR models were originally developed to model individual pathogens. More recently, these models were extended to multiple compartments, which allows the modeling the dynamics of multiple disease strains [3–6] and cross-immunity in an age structured model [7–9].

A different approach to modeling disease spread is a data-driven approach, using the most recent known disease load and/or proxy data to disease load. The former is usually highly accurate but is gathered with some delay and at lower temporal and spatial resolution, compared to the latter, which is less accurate but can be collected in near real time with relatively high spatial resolution. Examples of such proxy data have, in recent times, focused on Internet data such as search queries [10–13], social media postings [14,15] or other Internet data [16]. Social networks often reflect dynamics of aggregation of types and subtypes of influenza, and commonly serve as crude estimation to the real number of infected individuals. Monitoring influenza by type and subtype not only provides more detailed observational content but supports more accurate forecasting [17].

However, when tracking disease with such proxy data it is usually impossible to distinguish between similar diseases because data are usually related to symptoms of the disease. In the case of influenza, these include low specificity symptoms such as cough, sore throat, and fever of $\geq$37.8 °C. Hence, tracking of influenza is usually replaced by tracking of influenza-like illnesses (ILI). ILI includes diseases with similar symptoms, such as respiratory syncytial virus (RSV) and parainfluenza. If (as is usually the case) several diseases overlap in time and space, ILI rates cannot be modeled by simple SIR models, and the important link between data-driven disease tracking and tracking via epidemiological models is lost.

Prior research has attempted to bridge this gap. However, simple application of multiple SIR models is limited by the dimensionality of the resulting problem. Modeling of n viral strains in a simple history-based model requires the solution to $O(2^n)$ equations, making it difficult to include more than a few strains. As a result, most previous work using a population level model has either focused on a small number of strains [18,3,19], or reduced the dimensionality of the model by making certain strategic assumptions [20,21]. In particular, it is possible to reduce the history-based equation system to $O(n)$ equations if all strains are tracked, but the order in which they are seen is not [22,23]. This limits the number of potential strains that can be studied unless computationally intensive individual-based models are used [24,25].

Thus, here we focus on modeling multiple pathogens using an ensemble of SIR models. We show that the temporal dynamics of ILI can be blindly partitioned into multiple SIR models in a data-driven manner and show that the resulting models correspond to known pathogens.

2. Modeling

Our modeling approach is based on the basic SIR model, due to Kermack–McKendrick [26]. It describes the evolution of an infectious agent in a population using a system of ordinary differential equations:

$$\frac{dS}{dt} = -\beta SI, \quad \frac{dI}{dt} = \beta SI - \gamma I, \quad \frac{dR}{dt} = \gamma I \tag{1}$$

where $\beta > 0$ is the infection rate and $\gamma > 0$ is the recovery rate. To improve this model and account for multiple viruses infecting the same population we suggest to transform (1) to be multidimensional, thus representing the dynamics of a population infected by multiple viruses and virus strains. To this end, we assume a v dimensional space, where v denotes the number of distinct viruses existing in a specific population prone to infection. If this is the case, S, I, and R take the form of square matrices with dimension v. The specific elements in these matrices, e.g. S_{ij}, are defined for viruses of type i in case that $i = j$, or when $i \neq j$, as a mutation of virus i to virus j. Note that only pairwise mutations are possible in the matrix form of the equation. Higher order mutations will require tensor representation thereof.

We force 0 to specific elements in S, I, and R when the option to mutate does not exist or its likelihood is negligible. Specifically, in this work we assume that mutation is unlikely. Therefore, all non-diagonal elements of the matrices (e.g., $i \neq j$) are equal to zero. However, as we argue in the Discussion, the non-diagonal case can be important for modeling multiple geographic areas or different demographic groups.

β in its matrix form accounts for the infection rate and the ability to cross infect. Similarly, γ is defined such that in the case of $i \neq j$ we account for a different recovery period for people who are infected by a mutation and were infected in the past by a similar virus.

In the general case the new dynamics can be defined as:

$$\begin{aligned} \frac{d\mathbf{S}}{dt} &= -\mathbf{I}\boldsymbol{\beta}^{\top}\mathbf{S} \\ \frac{d\mathbf{I}}{dt} &= \mathbf{S}\boldsymbol{\beta}^{\top}\mathbf{I} - \boldsymbol{\gamma}^{\top}\mathbf{I} \\ \frac{d\mathbf{R}}{dt} &= \boldsymbol{\gamma}^{\top}\mathbf{I} \end{aligned} \tag{2}$$