Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Annals of Epidemiology 65 (2022) 1–14

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

Original article

Mitigating bias in estimating epidemic severity due to heterogeneity of epidemic onset and data aggregation

R.G. Krishnan[a], S. Cenci[a,b], L. Bourouiba[a,c,*]

[a] Massachusetts Institute of Technology, Cambridge, MA
[b] Imperial College London, UK
[c] Health Sciences & Technology Program, Harvard Medical School, Boston, MA

ARTICLE INFO

Article history:
Received 9 December 2020
Revised 11 June 2021
Accepted 18 July 2021
Available online 19 August 2021

Keywords:
Epidemiology
Reproduction number
Aggregation bias
Influenza
Delays in epidemics
COVID-19

ABSTRACT

Outbreaks of infectious diseases, such as influenza, are a major societal burden. Mitigation policies during an outbreak or pandemic are guided by the analysis of data of ongoing or preceding epidemics. The reproduction number, R_0, defined as the expected number of secondary infections arising from a single individual in a population of susceptibles is critical to epidemiology. For typical compartmental models such as the Susceptible-Infected-Recovered (SIR) R_0 represents the severity of an epidemic. It is an estimate of the early-stage growth rate of an epidemic and is an important threshold parameter used to gain insights into the spread or decay of an outbreak. Models typically use incidence counts as indicators of cases within a single large population; however, epidemic data are the result of a hierarchical aggregation, where incidence counts from spatially separated monitoring sites (or sub-regions) are pooled and used to infer R_0. Is this aggregation approach valid when the epidemic has different dynamics across the regions monitored? We characterize bias in the estimation of R_0 from a merged data set when the epidemics of the sub-regions, used in the merger, exhibit delays in onset. We propose a method to mitigate this bias, and study its efficacy on synthetic data as well as real-world influenza and COVID-19 data.

Introduction

The burden of infectious diseases is felt around the world. Developed countries are often faced with increased health-care expenditure, while developing countries face increased mortality and detrimental long-term effects on the population [1]. Seasonal epidemics such as influenza infect millions of people annually since preventative measures are imperfect. In the 2017–2018 epidemic, there were 45 million influenza illnesses and 21 million influenza-as-sociated medical visits. In the 2018 – 2019 epidemic, the efficacy of the flu vaccine was reported to be as low as 47% [2]. Influenza severely affects vulnerable groups in particular, such as young children and the elderly [3], and in 2017, the associated death toll was around 61,000 in the United States (US) [4]. Moreover, the total economic burden of influenza epidemics within the US is projected to be in the realm of $87.1 billion [5]. The ongoing COVID-19 pandemic has affected more than 251 million people worldwide with a fatality of 5 million [6], and an economic contraction nearing that of the Great Depression of the 1930s and with projected injections of more than 5 trillions globally [7]. Economists estimate, under the optimistic assumption that the pandemic's effects will abate in two years, the total cost of the pandemic would lie at nearly 16 trillion dollars [8] and we are now seeing the emergence of new strains. Clearly, the severe human and economic costs of pandemics and regular seasonal epidemics motivate the need for accurate methods for risk-prediction and planning during outbreaks. To this end, Non Governmental Agencies (NGOs) and health care officials work in tandem to collect disease incidence data.

Epidemiological models can be used to turn the collected data into actionable insights [9–11]. The parameters in epidemiological models quantify the severity of the disease [12] and the parameters thus obtained are used to devise strategies of intervention [13–15]. The reproduction number, R_0, is a critical quantity defined as the expected number of secondary cases caused by a single infected individual in a population of susceptibles [16–18]. The reproduction number characterizes the growth potential of an epidemic [19]. It also serves as a threshold parameter: if above one, an epidemic is expected to grow, while if below one, an epidemic

* Corresponding author.
E-mail address: lbouro@mit.edu (L. Bourouiba).

https://doi.org/10.1016/j.annepidem.2021.07.008
1047-2797/

is expected to decay. R_0 is widely used for risk-assessment, optimization of strategies for intervention or management of allocation of human and economic resources, particularly during the first stages of a crisis [20–22]. An accurate estimation of the reproduction number from observational data is thus critical. However, such assessment relies on the accuracy of collected epidemiological data, which is often noisy due to a variety of factors which we shall discuss next.

As exemplified by the ongoing COVID-19 pandemic, the collection of data during an outbreak is challenging because of logistical, cultural, political, and technical factors, particularly at early stages of an epidemic or pandemic. For example, under-reporting of cases, and slow transmission of data between organizations were important sources of heterogeneity that made their way into the final estimates collected by the World Health Organization (WHO) during the Ebola outbreak [23]. In the case of influenza in the United States, data are collected in a hierarchical structure: incidence is recorded at sub-regional organizations such as local hospitals, agglomerated by regional organizations (such as state level healthcare agencies) before being transmitted to the Centers for Disease Control and Prevention (CDC) and aggregated to create national infection curves [24]. Modern multi-regional data collection and aggregation practices may no longer reflect the assumptions of typical epidemiological models. *Is such an aggregation approach valid when the epidemic has different dynamics across the regions monitored?* In this study, we characterize bias in the estimation of R_0 from a merged dataset when the epidemics of the sub-regions, used in the merger, exhibit delays in onset. We propose a method to mitigate this bias, and study its efficacy on synthetic data as well as real-world influenza and COVID-19 data.

The paper is organized as follows: in Section 2, we showcase a real-world example of variation in epidemic onset, and discuss how such variation arises in epidemiological data. Section 3 provides background on the SIR model and how its parameters are interpreted. In particular, we discuss how temporal delays in the onset of an epidemic affect the epidemic curves and bias estimates of the reproduction number. Section 4 details the method we propose to reduce this bias by explicitly accounting for temporal offsets, introducing a new Shifted-SIR model (S-SIR). In Sections 4 and 5, we validate our method using synthetic data and real-world influenza data collected in the United States, respectively. Finally, in Section 6 we use our methodology to illustrate how the methodology can help reduce errors at early stages of epidemics even when understanding of the underlying disease dynamics is not yet clear. We conclude with a discussion of related work and the implications of our method in Section 7.

Heterogeneity in epidemic data

The procedures for reporting cases of infectious diseases vary from country to country but generally follow a hierarchical structure. Starting from hospitals and clinics, data is collected, tabulated and aggregated before being passed onto regional authorities and NGOs. This process is repeated and the aggregated statistics are sent to national bodies where it is used for intervention policy, crisis response, planning and management. In resource-limited settings, some of this hierarchical infrastructure has to be created on-the-go during an outbreak [26].

By way of example, consider a government agency in the United States that wishes to quantify the severity of the influenza epidemic in 2014. Typically, the severity of an epidemic is assessed by estimating the R_0 from data collected locally and then aggregated.

However, the aggregation process smoothes out local heterogeneity that are relevant for the assessment of the severity of the epidemic. For example, in Fig. 1 we plot the incidence curves for the 2014 influenza outbreak in several major states in the United States; here we observe non-uniformity in the start times of the epidemic across the various (sub-regions) states. Week 40 of the year is typically used by the CDC to denote the start of the flu season [25].

Decisions on quantifying the severity of the disease are made using national level epidemic data and the aggregation infrastructure that creates the national data pools the data from the sub-regions. Typical compartmental models assume that incidence curves reflect a single large well-mixed population. This assumption is violated when sub-regional epidemics begin at different points in time. Failing to account for this heterogeneity in onset of the epidemic among sub-regions can bias estimates of R_0 inferred from the aggregated data. We pause to reflect on *how* such heterogeneity in onset time of epidemics could arise.

1. **Errors in Data Reporting:** An infection curve typically represents the number of infected individuals over time. However, the reported number is often a noisy estimate of the true number of individuals infected. Deviations from ground truth can occur due to under-reporting [27,28], or errors in tabulation.
2. **Ecological bias:** Typical compartmental models assume a *single*, mean, rate of infection and recovery across the entire population. Aggregating incidence counts from sub-regions in which there is high variance in the rates of infectivity and recovery may result in inferring a global reproduction number, R_0 [29] that is not representative of the reproduction numbers of sub-regions. This is a case where ecological bias can arise in epidemiological studies [30]. Correcting ecological bias of this nature is difficult due to non-identifiability in the data: namely that there exists an infinite set of sub-regional incidence curves, each corresponding to different values of R_0, that can give rise to an observed aggregated infection curve.
3. **Offset in epidemic onset in sub-regions:** Epidemics rarely begin simultaneously across spatially separated geographic regions. Due to factors such as the population density of sub-regions and the variation in disease vectors, there are often delays in the onset of the epidemic. These delays can be difficult to detect: for example, California in Fig. 1, has a non-zero number of cases of influenza as baseline *before* the onset of the epidemic, making it difficult to pinpoint exactly how to define the onset of the epidemic. Organizations like the CDC often use week 40 in the year as an empirical mark of the beginning of seasonal epidemics such as the flu. This is a choice that aligns well with the epidemic dynamics in some sub-regions but not others.

We next focus on mitigating errors in estimating R_0, and thereby the quantification of epidemic severity, due to the third scenario highlighted above. We expect that when delays in epidemic onset among sub-regions are comparable to the duration of the epidemic, the aggregated curve is fundamentally altered. We conjecture that the resulting bias in R_0 arises because parameters in compartmental models are forced to capture delays, in addition to the dynamics of the epidemic.

Hereafter, sub-regions refer to collection sites for epidemic data while regional data refers to the aggregation of incidence counts from such sub-regions to form the incidence curve from which R_0 is inferred. We work with the canonical epidemic compartmental model, the SIR model, described succinctly in the next section.

The SIR model

Form and basic assumptions

The Susceptible Infectious Recovered (SIR) model is ubiquitous for the analysis of epidemic data [31]. It splits a homogeneous population into three groups, or compartments, and represents the

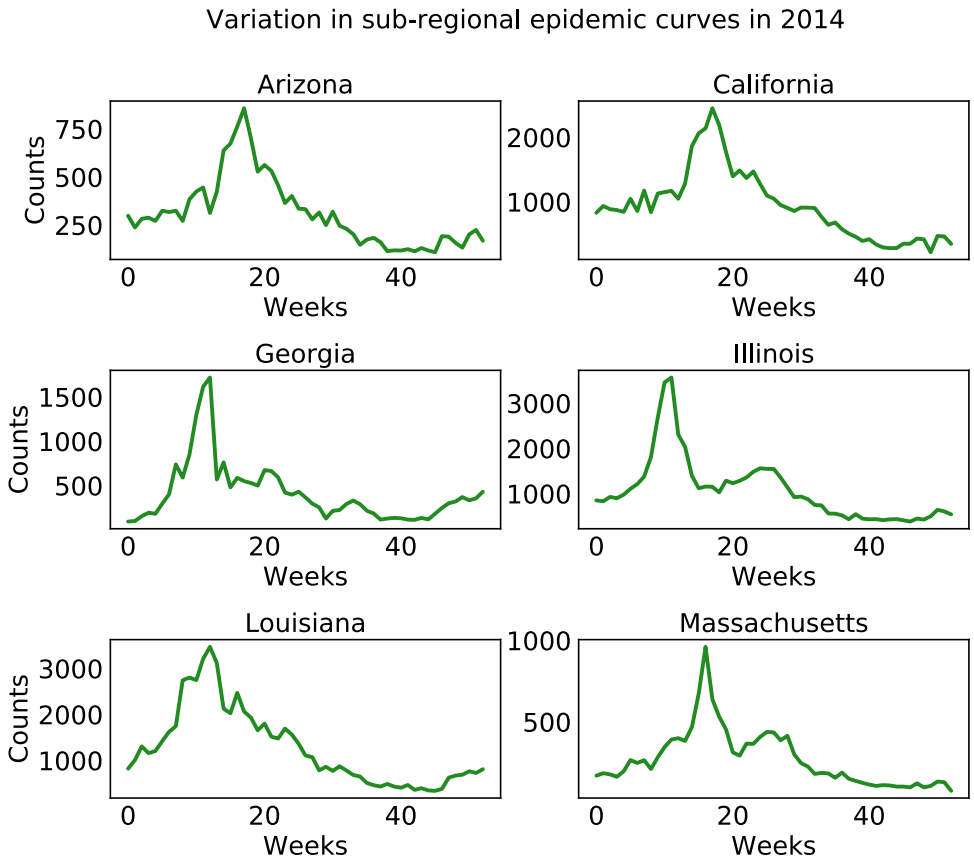

Fig. 1. Flu outbreaks in 2014: Infection curves for influenza across different states of the U.S. On the x-axis is weeks elapsed since week 40 [25], used by the CDC as a marker for the onset of the flu season. On the y-axis is the number of individuals who report to outpatient clinics with an influenza like illness. Note that the maximum number of infected individuals differs across sub-regions.

progression of the epidemic as the status of individuals transition through the compartments. Typically, nearly all individuals are initially considered to be Susceptible while a small number of infected individuals are introduced in the naive population. The rates of change of the proportion of individuals in each compartment are governed by coupled ordinary differential equations governing the number, or fraction, of the susceptible (S), infected (I) and recovered (R) individuals:

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

We work with the normalized model where $S + I + R = N = 1$. β is the mass action parameter encompassing contact and transmission rates between susceptible and infected individuals, and γ, the rate of recovery of infected individuals.

The reproduction number, the average number of secondary infections caused by a single infected individual introduced in the population, is derived as $R_0 = \frac{\beta S_0}{\gamma}$ where S_0 is the initial fraction of susceptible individuals in the population.

The primary use of R_0 is as a threshold parameter to quantify whether an outbreak is expected to die off or spread and grow into an epidemic. When $R_0 > 1$, the region experiences an epidemic with an increase in the number of cases, I. When $R_0 < 1$, the outbreak dies out (I decreases). The value of R_0 is therefore crucial for risk assessment and testing intervention strategies such as planning for the number of drugs to allocate, hospital beds to prepare, and healthcare workers to deploy during the epidemic. An under or over-estimation of the reproduction number can have disastrous public health and economic consequences.

Note that the SIR model makes a number of implicit assumptions such as homogeneity in the distribution of the recovery and infection rates. Anyone is equally likely to be infected and recover. The model also assumes homogeneity in the social contact network: i.e. any infected individual is equally likely to be the source of infection for any other individual, not accounting for distance, heterogeneity in pathogen shedding, residence in the same indoor space, or contributions from air contamination indoors for respiratory diseases for example [32–34]. Several extensions to the model have been proposed to account in various ways for homogeneity in spatial and temporal scales or shedding amounts and timescale competitions [35–39], but there is no universal approach yet to best account for these effects in general.

We begin with a demonstration on synthetic data to show how delays in onset of epidemics in sub-regions can affect assessment of R_0 from the regional data.

Motivating the problem with the SIR model

The setup is as follows. We consider an epidemic with $R_0 = 2.3$ as manifested in ten different sub-regions. Each sub-regional epidemic is characterized by a choice of $R_0 \sim |\mathcal{N}(2.3; 1)|$. The incidence curves are simulated and visualized in Fig. 2 [**A**]. The regional data is given by their aggregation in Fig. 2 [**B**]. Although typically, R_0 would only be estimated from regional data, in this exercise, we use the SIR model to infer R_0 from each of the sub-regional curves and their aggregation – we visualize the results in Fig. 2 [**C**]. Next, we repeat the exercise in Fig. 2 [**D-F**] but this time simulate epidemics where the epidemic onset in the sub-regions is delayed, with a delay uniformly sampled from two to six weeks.

Without delays in onset of epidemics in sub-regions, even with noise in the observational data and a small degree of variation in

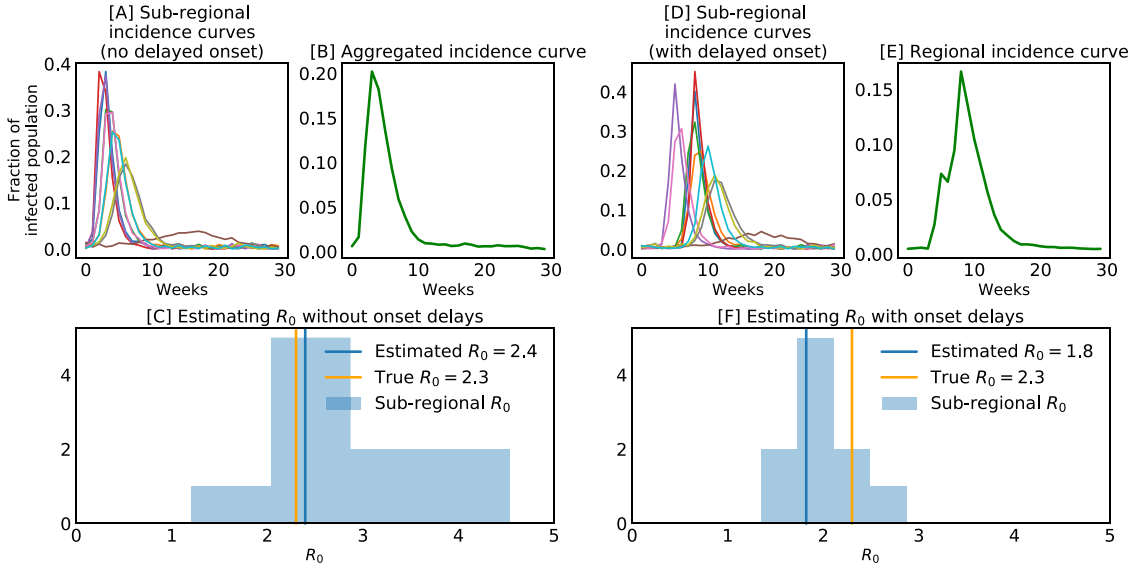

Fig. 2. Time-delay is a source of error when inferring R_0. [**A**] shows the infection curves for populations in ten hypothetical regions alongside the aggregated infection curve in [**B**]. [**C**] shows the results of inferring R_0 using the SIR model on all the sub-regional data, the aggregated curve along with the *true* $R_0 = 2.3$. [**D**], [**E**], and [**F**] follow similarly but showcase instances where there is heterogeneity in epidemic onset time between the sub-regions with resulting values of R_0 estimated using the SIR model that are much further away from the ground truth.

the sub-regional R_0, inferring R_0 from the aggregated data using the SIR model yields a reasonable estimate $R_0 = 2.4$ compared to the true value of 2.3. However, in the presence of delayed epidemic onsets in the sub-regions, the inferred $R_0 = 1.8$ from regional data is less accurate, further from the true value. This demonstrates how a naive approach to the inference of the reproduction number can lead to a serious underestimation of the severity of an outbreak.

To illustrate the importance of the problem, let us consider a population of a million individuals. In the absence of temporal offsets, the model's estimation, $R_0 = 2.4$, (blue line in panel [**C**]) would be approximately 240,000 infected individuals while the true number, $R_0 = 2.3$, (orange line in panel [**C**]) predicts 230,000. This is an under-estimation of 10,000 individuals or 1% of the population. In the presence of heterogeneity due to variation in epidemic onset (panel [**F**]), the number of *underestimated* infected individuals rises up to 200,000. This is 20% of the population, and four times larger than the error incurred in the absence of temporal variation in disease onset. Such large errors in the prediction of the number of infected individuals can have dramatic consequences during an outbreak; for example: from an operational standpoint, care facilities may be under-prepared for dealing with the additional influx of sick patients.

Note that in the experiments discussed here, we ensured overlap between the sub-regional epidemic curves in Fig. 2 [D]. Clearly, in the extreme case in which the epidemic curves do not overlap, the use of an SIR model for the aggregated model can no longer hold. The aggregated curve would either be in the form of a plateau or appear as a series of epidemic curve peaks, possibly suggesting multiple epidemic waves. These extreme cases imply obvious breakdowns of the SIR type model for the aggregated data, and so are *not* of high concern: a typical user would not typically try to extract SIR-type parameters from such multi-modal or "flat" epidemic data curves. However, more concerning is the intermediate regime for which delays in onset are important, but with sufficient remaining overlap between epidemic curves, such as shown in Fig. 2 [D]. In such cases, there is no obvious distortion of the aggregated epidemic curve. As we have seen, in such cases, the R_0 inferred is typically biased and in particular, is underestimated. Indeed, in such intermediate cases, the SIR model is incapable of capturing time-delays, or offsets, in the onset of the epidemic in sub-regions, the source of errors we seek to avoid. To remedy this pathology, we next introduce an alternative epidemic model.

Shifted-SIR (S-SIR) model

The shifted-SIR model uses an additional parameter to explicitly capture delays in the onset of the epidemic.

Denoting by τ the start time of the epidemic, and $\mathbb{I}$ the indicator function:

$$\mathbb{I}[x] = 1 \text{ if } x \text{ holds, } 0 \text{ otherwise.}$$

The S-SIR model has the following dynamics for each of the compartments of a population:

$$\begin{aligned} \frac{dS}{dt} &= \mathbb{I}[t > \tau](-\beta SI), \\ \frac{dI}{dt} &= \mathbb{I}[t > \tau](\beta SI - \gamma I), \\ \frac{dR}{dt} &= \mathbb{I}[t > \tau](\gamma I). \end{aligned} \tag{2}$$

The key difference between Equations (1) and (2) is the incorporation of an additional parameter to the model, τ. When $t > \tau$, the Shifted SIR, and the SIR model are equivalent. When $t \leq \tau$, the shifted SIR model preserves the initial conditions of the differential equation without temporal evolution. In this way, the model has a mechanism to explicitly account for the start time of each infection curve, and the new parameter has a rooting in a physical, measurable effect of delay in onset or reporting, or a combination of both. Consequently, this means that the parameters β, γ that govern the infection dynamics, and by proxy reproduction number R_0, are not used to model the early portion epidemic data attributed to delays and noise; they can instead focus on modeling infection counts after τ time-steps.

Simulation and parameter estimation in the S-SIR model: For a given choice of τ, one can simulate from this model using an SIR model followed by right-shifting the simulated curve right by τ time-steps. However, the incorporation of τ to the compartmental model requires new methods to fit parameters from epidemiological data. We use a grid search for τ and nested within it, a least-

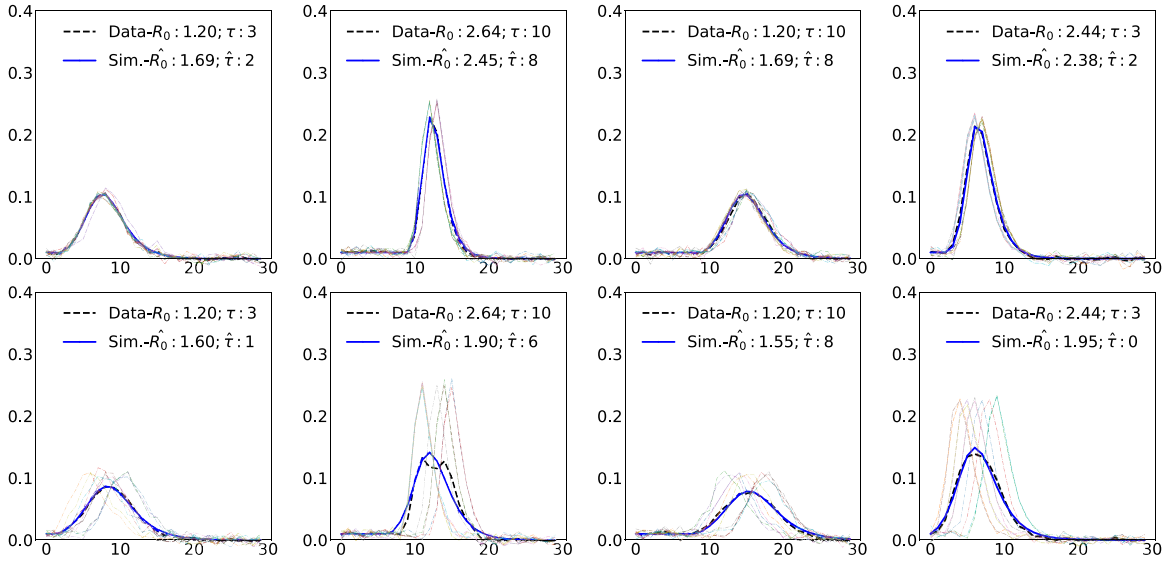

Fig. 3. Inferring parameters of the S-SIR model For each sub-region, we simulate noisy data following SIR dynamics with severity R_0 and time-delays Uniform$(\tau-1, \tau+1)$ (top row) and Uniform$(\tau-3, \tau+3)$ (bottom row) using the values of R_0, τ displayed in the top row of the legend. The aggregated curve (*Data*) is then used to infer the parameters of an S-SIR model and we simulate from the resulting model (*Sim.*; alongside the inferred parameters). The faded graphs are the epidemic curves in the sub-regions.

squares [59–61] based fitting procedure to estimate β, γ from infection counts I. The practitioner's prior estimates for what values τ can take for a given epidemic may be used to constrain the grid search. We defer the reader to the supplementary information (SI) for a detailed exposition on how the parameters of the model are inferred.

Inferring parameters of the S-SIR model: We first reflect upon how heterogeneity in epidemic onset manifests in aggregated incidence curves. In Fig. 3 we study eight different scenarios, where both R_0, τ take high and low values (values listed in the figure-key labelled as *Data*). In the top row, each of the ten sub-regional epidemic curves (faded curves) are assumed to be generated from an SIR model with the corresponding value of R_0 and shifted to the right by an offset drawn from Uniform$(\tau-1, \tau+1)$. In the bottom row, each of the ten sub-regional epidemic curves (generated as above) are shifted to the right by an offset drawn from Uniform$(\tau-3, \tau+3)$. The top row illustrates the case where there is a small degree of heterogeneity in onset times among the sub-regions and the bottom row illustrates the case where there is a large degree of heterogeneity. The aggregated curve (depicted with the label *Data*) is then used to infer the parameters of the S-SIR model. We simulate from the model and display the estimated parameter values $(\hat{R_0}, \hat{\tau})$ alongside the curve from the simulation, noted with *Sim* in the figure-key.

The results showcase a few important aspects of the inferred parameters. First, across the board, we find that simulation from the S-SIR model with the inferred parameters presents an accurate fit to the aggregated, epidemic curve. Second, when the epidemic onset delays in the underlying sub-regions have a small amount of variation (top row), when they differ greatly (bottom row) *and* the epidemics among the sub-regions have the same fraction of infected individuals, the value taken on by $\hat{\tau}$ lies close to the smallest time-delay among the sub-regions. We conjecture that this is because the smallest time-delay among the sub-regions is the minimum delay that may be identified from the aggregated infection counts. In practice the degree of observational noise, the number of sub-regions, the severity of the epidemic in each sub-region and the heterogeneity in delay onset affect the values of $\hat{R_0}, \hat{\tau}$ inferred.

Improving R_0using the S-SIR model Next, we perform a more thorough quantitative study of scenarios to understand where the S-SIR model can improve the estimation of R_0 in the presence of offsets in epidemic onset in sub-regions.

The setup for the experiments are as follows. Across either ten or one hundred sub-regions, we first simulate an SIR epidemic and shift the epidemic curve to mimic a delayed onset. Folded random noise (with standard deviation of 0.005) is added to each point in every simulated curve. The curves are aggregated and the aggregation is used to infer the parameters of the SIR and the S-SIR, using $\mathbb{O}(\mathbb{X}, 0)$ and Algorithm 1, respectively. We then study the error between the inferred value $\hat{R}$ and the true value, R_0. Each such experiment is repeated 1000 times.

We study two kinds of epidemics with different severities: $R_0 = 1.8$, $R_0 = 3.1$. We vary the way in which the delay in epidemic onset in sub-regions, τ, behaves among the sub-regions by selecting four distributions from which τ (in unit of weeks) is sampled:

1. $\tau \sim \Gamma(0.3)$
2. $\tau \sim \Gamma(5)$
3. $\tau \sim$ Uniform$[0, 3]$
4. $\tau \sim$ Uniform$[7, 10]$

The results of our analysis are depicted in Fig. 4 when τ is drawn from a Gamma distribution among the sub-regions and in Fig. 5 when τ is drawn from a Uniform distribution. Both violin plots indicate that the SIR and S-SIR models *underestimate* R_0. However, across all results, we find that errors in the estimation of R_0 inferred from the S-SIR model are on average, lower than errors in R_0 from the SIR model, suggesting that explicitly accounting for time-delays does improve the accuracy of the inferred R_0. We examine several follow up questions in turn.

Accuracy in the absence of time-delays:

Figs. 4 and 5 show the results of the inference with small time delays (two columns on the left of all the plots of the figures). We note that from the perspective of planning for epidemics, it is worrying that even small time-delays in epidemic onset in sub-regions result in an appreciable error in the estimation of R_0 from aggregated data, giving a lower value than reality on the ground; this showcases a scenario that calls for concern about bias in the estimates of R_0. Such scenario is particularly concerning when the epidemic is severe (relatively high R_0). The Figs. (4 and 5) show that both synthetic scenarios tested, we find some gains from the

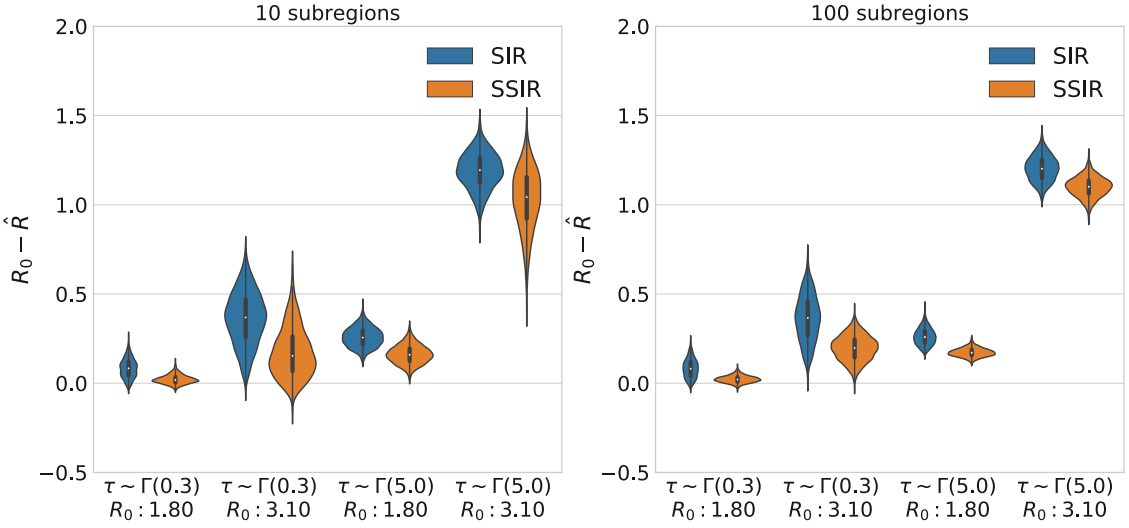

Fig. 4. Synthetic results (Gamma distribution): The number of sub-regions on the left plot is 10 and on the right is 100. In each, we study the distribution of errors between the true R_0 and inferred values when the time-delays and severity of infection with the sub-regions are varied. We find that the S-SIR systematically obtains better results across all settings.

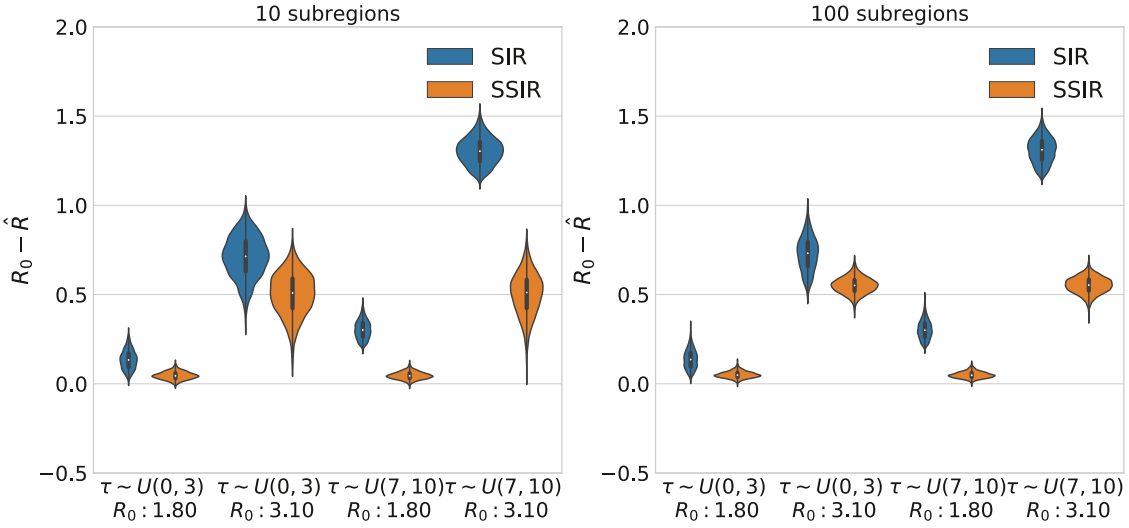

Fig. 5. Synthetic results (Uniform distribution): The number of sub-regions on the left plot is 10 and on the right is 100. In each, we study the distribution of errors between the true R_0 and inferred values when the time-delays and severity of infection within subregions are varied. We find that the S-SIR systematically obtains better results across all settings but does particularly well when there are larger time-delays in the sub-regions.

use of the S-SIR model which has a lower (on average) error in the estimation of R_0 than the SIR model.

Accuracy in the presence of time-delays:

When time-delays in sub-regions are larger (right two columns in the subplots of Figs. 4 and 5, we continue to find that the S-SIR model outperforms the SIR model. The improvements yielded by the S-SIR model are particularly visible when all the sub-regions have consistent shifts in the start of the epidemic and when the epidemic is more severe.

Over or under-estimation:

Across all the results we find that there can be a degree of under-estimation of R_0 when the sub-regions have time-delays, regardless of the method used. This is problematic; an underestimation of the epidemic severity can result in worse consequences on affected population than an overestimation of the severity.

The results on synthetic data demonstrate that the S-SIR model does improve the estimation of R_0 realizing values closer to the ground truth. The results also characterize scenarios when we can expect the use of such models to be of benefit: namely, when the delays in onset are large across sub-regions. We next discuss the insights gained from the synthetic data on characterizing the severity of epidemics from real-world influenza data.

Quantifying the severity of seasonal influenza

Influenza is a respiratory, viral, infectious disease which spreads via air. Seasonal influenza is recurrent; the young and the aged are particularly vulnerable to developing serious complications. Characterizing the severity of an epidemic both when it is ongoing as well as *a posteriori* is vital for public health management. We study how offset in epidemic onset times in aggregated data gathered from subregions in the United States affects the estimation of the severity of epidemics nationally. Epidemic data is collected in a hierarchical structure that spans multiple levels. Here, we focus on two levels of the hierarchy. The sub-regions will contain incidence counts available at the state level. The regional incidence counts correspond to those at the national level.

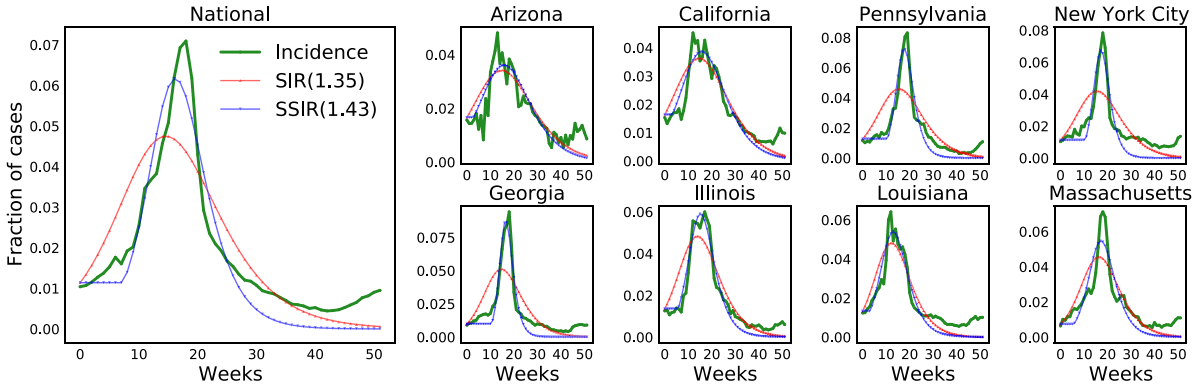

Fig. 6. Influenza outbreaks of 2017: On the right, across different states, each (green) curve represents the fraction of people in the infected (I) compartment. On the left is the aggregated National level infection curve. The comparison of the abilities of the SIR and S-SIR models to capture the influenza incidence curves is shown.

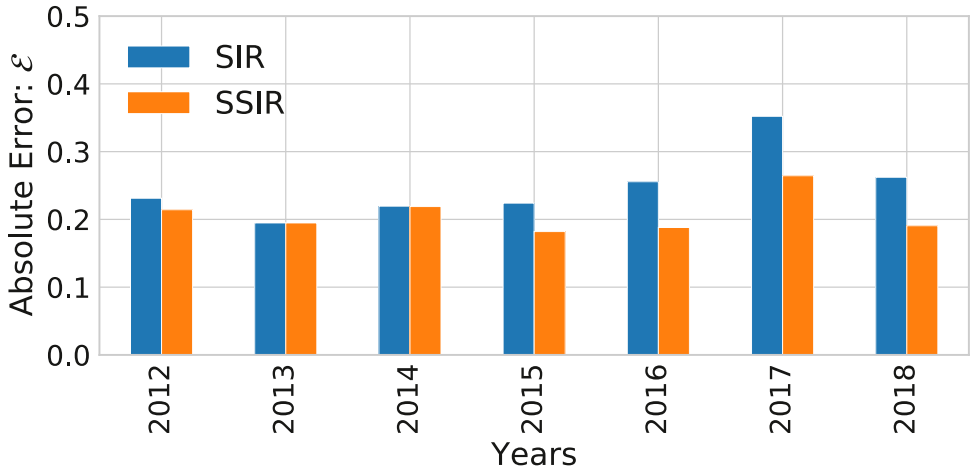

Fig. 7. Error on national – aggregated – level infection data: Across all the years we find that the S-SIR model incurs a lower error than that of the SIR model.

Setup: We obtained incidence reports for influenza from the CDC [40]. The data contains dates and raw counts of individuals diagnosed with influenza-like-illness (ILI) each week. For each state and year, we define time zero using the CDC's classification of the beginning of the flu year, which is the 40th week in the calendar year [25]. We self-normalize the raw counts (for example in Fig. 1) to create incidence curves where each point represents the fraction of the population in the Infected (I) compartment. The influenza data represents the incidence of the disease. Here, this is the number of new cases reported each week. Compartmental models typically involve prevalence: the number of infected individuals a given time. The two do not always coincide but since recovery from influenza typically takes five to seven days, this duration maps well with the temporal granularity of the available weekly case data. Thus, we make the assumption in this work that the two are interchangeable and learn the parameters of the compartmental models herein using the incidence data, similar to prior studies of influenza [41].

The first value of the incidence curve is used as the initial condition of Equations (1) and (2). For the S-SIR model, the $\tau_{\max}$ was limited to 20 weeks.

For this data, we no longer have access to the ground truth values of R_0. Therefore, as one proxy to measure how well the resulting model explains the data, we measure absolute error between the raw data and the curves simulated under each model with the learned parameters. Denoting I_s as the (vector of) data simulated from a model with learned parameters and I_d as the raw incidence data, we denote the absolute error $\mathcal{E}$:

$$\mathcal{E} = \frac{1}{T}\sum_{t=1}^{T} |I_s^t - I_d^t|. \tag{3}$$

A smaller absolute error implies that the parameters fit by the model more closely align with the observed data.

The effects of aggregation: Are there delays in sub-regional epidemics for influenza? We visually inspect the incidence curves among sub-regions in Fig. 6 to answer this question. We find a few instances where the epidemic onset is delayed. For example in populous states such as Illinois, California and Georgia, there is a large and sudden spike in the number of infected individuals over a short two or three week window. We can therefore suspect that the national aggregated level epidemic is affected as well.

In Fig. 6, we compare the results from SIR and S-SIR models learned on the incidence data both at the state level and at the aggregated national level. Here we verify that (a) the S-SIR model makes use of the parameter τ and (b) that doing so enables it to capture more accurately capture the peaks of the epidemic the national data and the sub-regional data. Both of these observations suggest that delays in sub-regional epidemics do play a role in how well R_0 is inferred on aggregated data. We conjecture this happens because the underlying epidemic data exhibits heterogeneity in the start of the epidemic among the sub-regions it is aggregated from (relative to the CDC's fixed point of observation for the start of the epidemic season). When the SIR model is fit directly to such data,

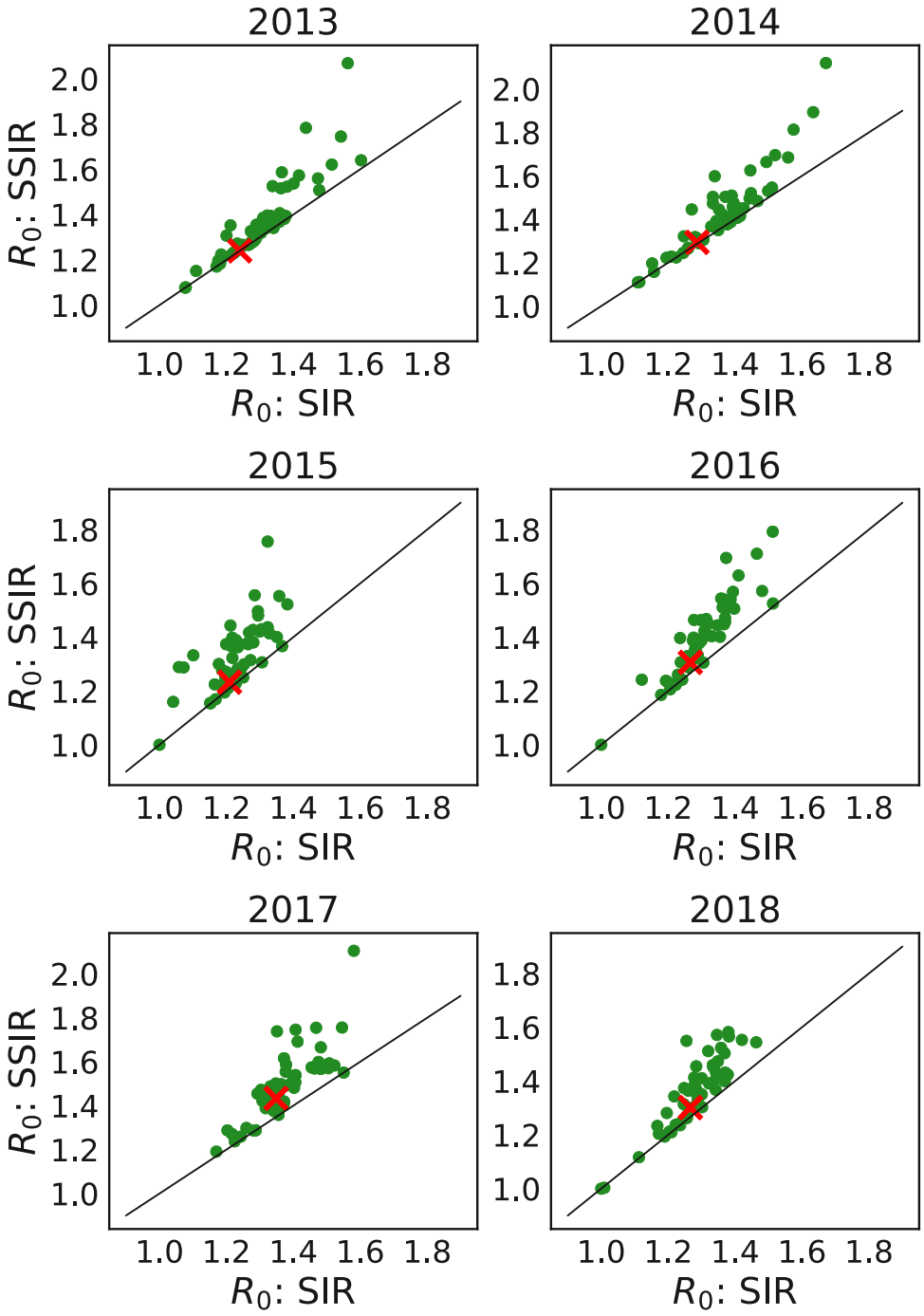

Fig. 8. R_0 **estimated from regional and national Incidence curves:** For six years, we visualize the values of R_0 estimated from regional (green) and national (red) levels incidence curves. The data was extracted and processed based on the incidence counts released by the CDC [42]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

it uses β, γ to fit early trends in the epidemic curve leaving it unable to capture the peak of the epidemic well. In contrast the S-SIR does not use β, γ to model the data until τ time has elapsed giving it the ability to model the peak of the epidemic curve more accurately.

Quantifying model performance: In Fig. 7, we quantify the absolute error obtained on national incidence curves for years ranging from 2012 to 2018. Across many of the years, we find that the S-SIR model has a lower error than the SIR model. In the supplemental material, we repeat this exercise at the sub-regional level – averaging the results across years from 2012 to 2018 – where we continue to find that the S-SIR captures the data better than the SIR model.

Comparing estimated reproduction numbers: We visualize the values of R_0 estimated by both algorithms in Fig. 8. On the x-axis are the results from the SIR model and on the y-axis are the results from the S-SIR model. In green are the (paired) estimates from each sub-region while in red are the estimates obtained from the national level incidence curves. We see a large degree of correlation between the values of R_0 inferred by the two algorithms. Across several years however, the value of R_0 estimated using the SIR model is always smaller than that estimated with the S-SIR

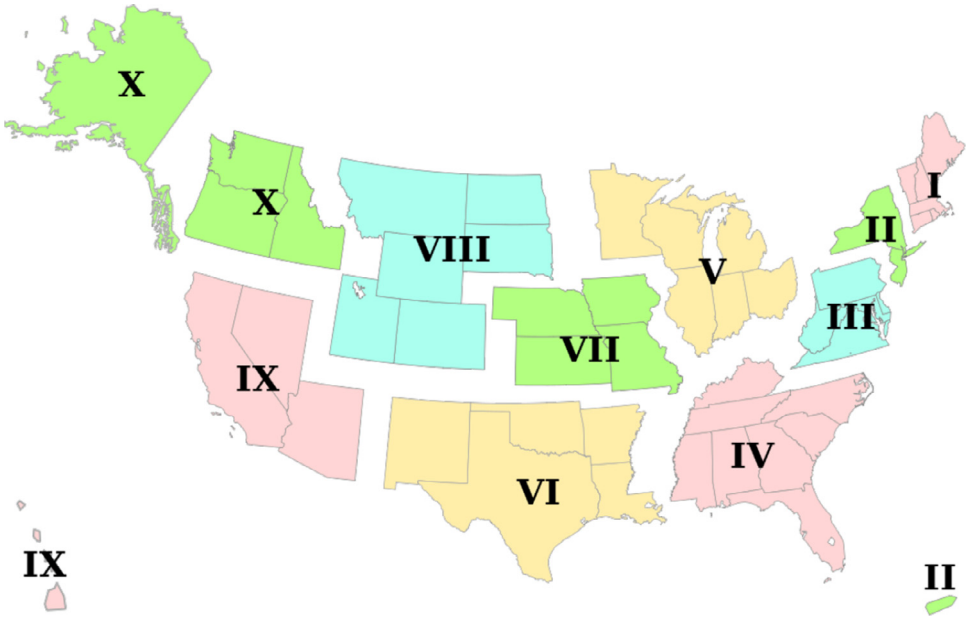

Fig. 9. United States Federal Regions [43].

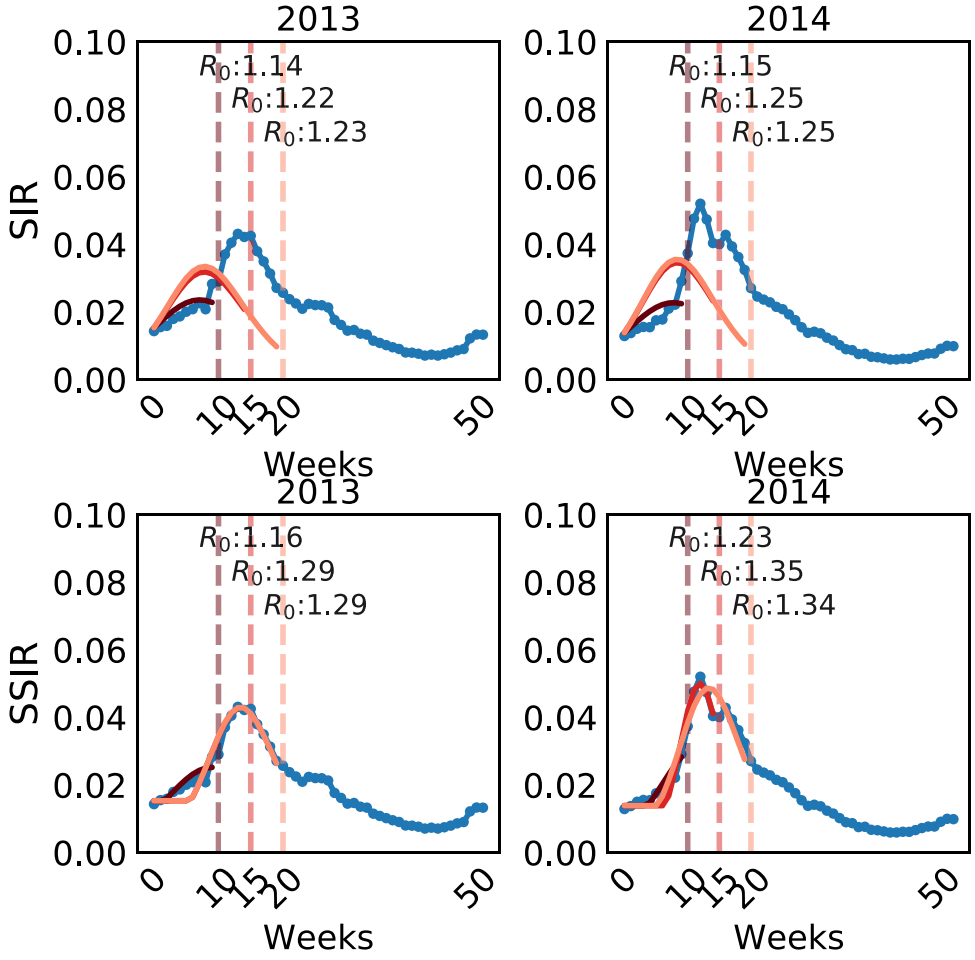

Fig. 10. R_0 **estimated from ongoing epidemics:** For two years, 2013 and 2014, we visualize the epidemic curve corresponding to the values of R_0 estimated from each year's national level incidence curve while limiting the length of the curve to 10, 15, 20 weeks. In each case, we visualize the results from the SIR and S-SIR models to infer R_0 and assess their model fit.

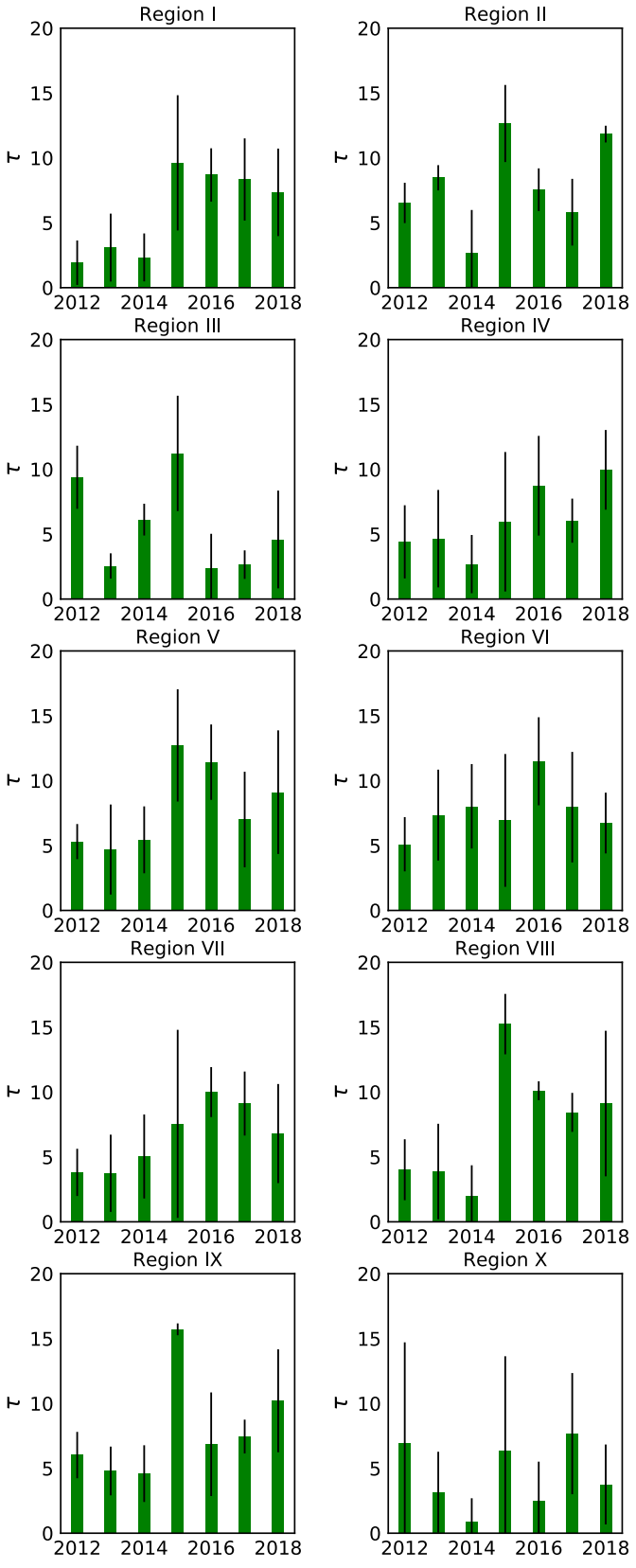

Fig. 11. Visualizing τ(unit of week) inferred by the S-SIR model across sub-regions (states): States are grouped based on the United States Federal Regions shown in Fig. 9.

model. Since we do not know the ground truth, it is impossible to know for certain which model is more accurate; however, the observation that the S-SIR model provides a better fit to the epidemic curve data (Fig. 7) along with Fig. 8, suggests that the SIR model underestimates R_0. We provide further evidence to support this hypothesis by looking at Fig. 12 where we visualized the learned τ from the S-SIR model. We see that years in which the S-SIR model performs better than the SIR model in Fig. 7 correspond to years where the inferred value of τ is largest.

Infectivity in ongoing epidemics: In the previous experiments, our analysis of the data was retrospective – i.e. the epidemic had come and gone. Using retrospective data, we can simulate an ongoing epidemic by limiting the length of the infection curve used to estimate the model parameters. For the national incidence curves for four years, we estimate the parameters of the SIR and S-SIR model using data up to weeks 10, 15, 20. For each choice, we calculate the value of R_0 and display it in Fig. 10. The top row depicts the results obtained from the SIR model while the bottom row contains the results from the S-SIR model. There are two observations of note – first, the S-SIR model captures best the peak of the epidemic; and second, the values of R_0 obtained, even during estimation of an ongoing epidemic, are higher than those obtained by the SIR model. Both observations suggest again that the S-SIR model yield more conservative, higher, estimates for R_0.

Heterogeneity in τacross states: Having studied how the S-SIR improves estimates of R_0 from regional data, we study the spatial patterns in the inferred values of τ from sub-regional data. We group the sub-regions based on standard federal regions and average the values of τ for states in each group. We visualize the federal regions in Fig. 9 and the results in Fig. 11. Across both time and space, we find that the S-SIR model infers a non-zero value of τ suggesting that the underlying epidemic is indeed beginning later than week 40, with large variations from region to region, and with several regions showing an increase in epidemic onset time in the 2015–2018 window (e.g., regions I, V, VII, VIII, and X). This can result in delays in onset of the national epidemic curve, as witnessed in Figs. 6 and 12, which subsequently interferes with the accuracy of the estimation of R_0 from national aggregated data.

Robustness of τ When only given access to the regional epidemic data, how robust are our estimates of τ to observational noise in the data? To answer this question, we use the national level epidemic curves, perturb the raw counts using additive Gaussian noise (with standard deviation set to a varying percentage of the peak infection counts during that year) while restricting ourselves to valid epidemic curves by using the absolute value when the addition of noise results in negative counts. Then, we re-estimate τ over one-hundred random trials. We visualize the results in Fig. 12. We vary the percentage of the peak infection count between 1% and 5% both of which represent a significant amount of observational noise added to the epidemic curve. Note that the noise (which is a constant function of the peak infection count) results in larger variation during the early and late stages of the epidemic curve than in the middle.

This experiment serves to assess the feasibility of using τ to anticipate whether the model we propose can assess the existence of time-delays from regional epidemic data. When the observational noise lies around 1% of the peak epidemic curve, we find that the variation is distributed tightly around the value of τ estimated from the noise-free regional data, implying that the use of τ is within reasonable limits for a practitioner to assess whether or not time-delay is an important effect in the data. When the standard deviation of the noise is increased to around 5% of the peak infection counts, we continue to find that in over 50% of the trials, the value of τ inferred indicates that one can expect time-delays in the underlying data.

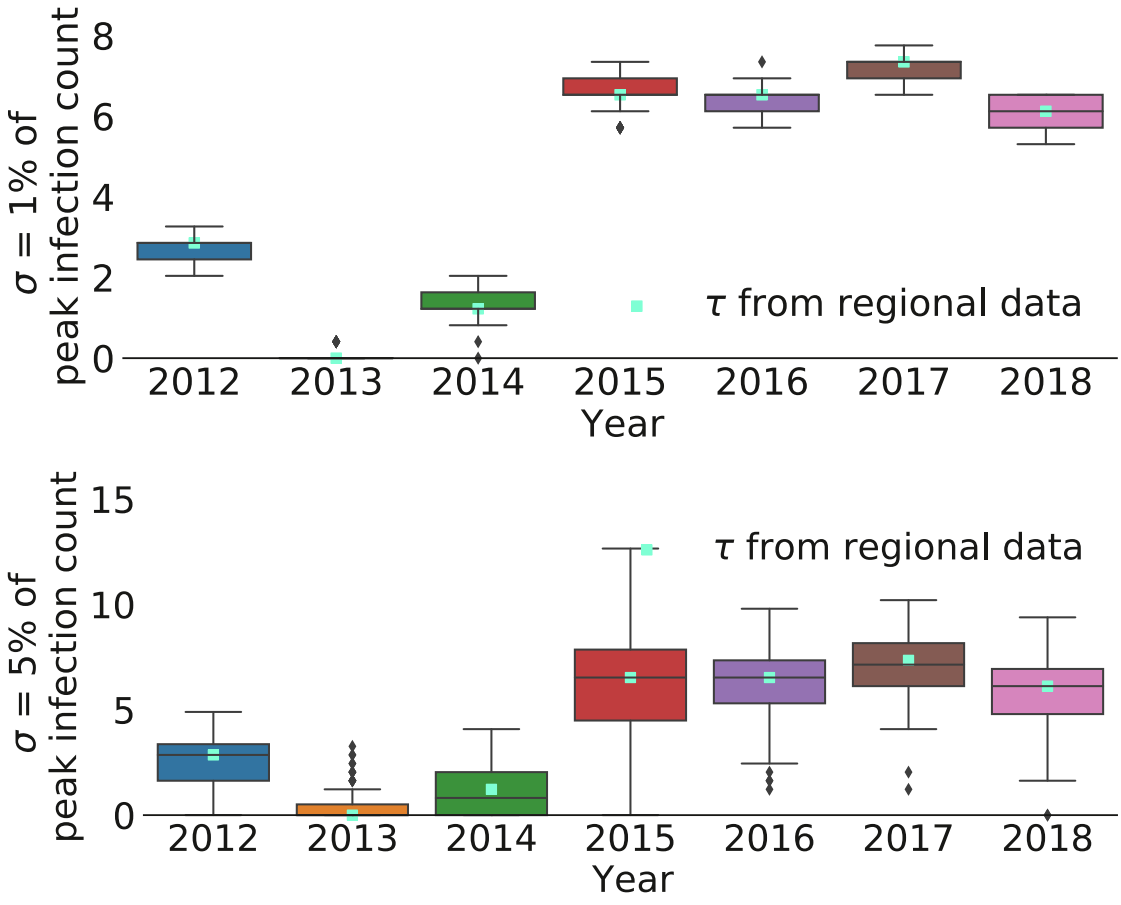

Fig. 12. τ **inferred from national (regional) level infection data:** Across all the years we show the τ from the S-SIR model. The intervals showcased are from re-estimating τ from epidemic curves with additive Gaussian noise .

Application to the ongoing COVID-19 pandemic

COVID-19 has spread over the globe, has infected over 251 million people worldwide, more than 5 million of whom have died [6]. Here, we compare the fit obtained to data from an S-SIR model with an SIR model on infection data collected from Italy at the early stage of the epidemic. It is important to note that by now, we know that the dynamics of COVID-19 is more complex than an SIR, but here we use only this approach to illustrate how the methodology can help reduce errors at early stages of epidemics even when understanding of the underlying disease dynamics is not yet clear. Given that the SIR model is the one initially typically used with new epidemics of unknown pathogens, we use an SIR rather than other models to make this illustrative point. Of course as knowledge of the disease evolution improve, we can use the same time-shifted approaches with more complex epidemic models, and this is beyond the scope of this manuscript.

The data we use for this illustration was updated daily, at the time, from the National civil protection department and can be downloaded from the official GitHub repository github.com/pcm-dpc/COVID-19.git. The data include information on those who were infected, recovered, and died. The epidemic began in Lombardy, a region in the north of Italy, and spread over the course of a month to the rest of the country. Fig. 13 (large panel) shows how R_0 changes as a function of the amount of data used to infer its value (on the x-axis) in the early stage of the epidemic. The error bars around the values of R_0 quantify the uncertainty on the estimate which is obtained by perturbing the infection curve with observational noise, and repeatedly estimating the reproduction number. The figure illustrates the significantly higher R_0 compared to that of influenza. The estimation is consistent with other, independent, estimates [44]. The figure also shows that the estimates start high, and decrease over time. The SIR and S-SIR estimates diverge in the early stages of the outbreak (large panel) but converge to consistent values further on (small panel). Furthermore, early estimate of R_0 from the S-SIR are significantly more robust than initial estimate from the SIR model. This suggest that the time-shift parameter of the S-SIR captures useful information in the early stages of an epidemic when data are scarce, prone to miss-reporting errors and when the underlying dynamics of the disease are not well understood.

Since the onset of COVID-19, our knowledge evolved and of course now we know that an SEIR may be more appropriate and that policy interventions over time have to also be accounted for to account for changes in the values of R_0 over time. Hence, we conjecture that in the early stage of the pandemic and onset of its various recurrent waves, the S-SIR could be a good model to estimate R_0, keeping in mind its limitations when complex interventions are being used.

Discussion

There is a large body of work studying various sources of heterogeneity that arise in the estimation of R_0 from data using epidemiological models. Extensions of the SIR model, such as the SEIR model [35], tackle heterogeneity due to differences in how diseases affect individuals in populations. Others focused on relaxing the assumptions made by typical compartmental models by incorporating variation in population sizes and the effect of vaccination [45] into the model of disease dynamics. In [46], heterogeneity in

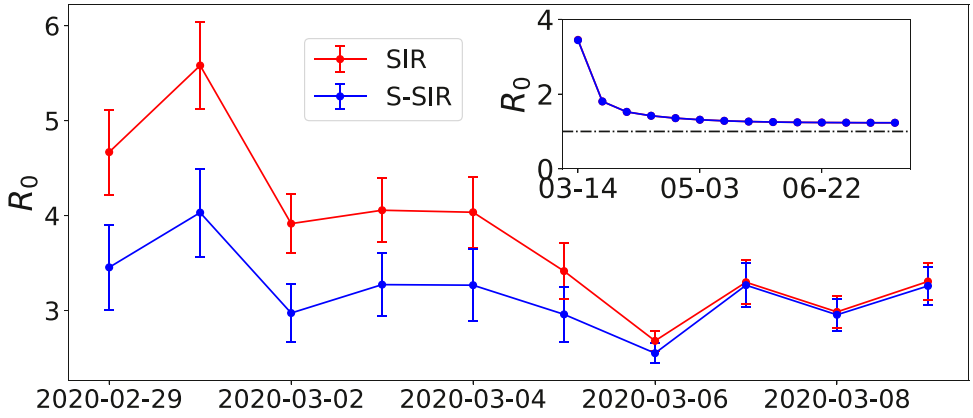

Fig. 13. R_0 **estimated from the 2020 COVID-19 pandemic:** the large panel shows the estimation of the R_0 using the national level curve data from the SIR versus S-SIR models in the early stage of the outbreak. The small inset panel shows the estimate in later stages. These estimates are not as reliable as in the later stages of the pandemic there were complex dynamics (e.g. policy intervention) in place.

removal from the *Infected* compartment is studied and its effect on R_0 is derived. [32] studies rates of infection in diseases where there is population heterogeneity (such as the level of social activity displayed by individuals), social-distancing and changes in hygiene. In the context of the Ebola outbreak in West Africa, [47] discusses the importance of accounting for uncertainty in the estimated model parameters. Variation in epidemiological curves and its parameters as a function of the variation in assumptions about the contact networks were also studied [48,49].

Biases incurred due to inferences made at the level of aggregate data is referred to as *aggregation bias* [50,51]. For example, [51] examined changes in regression coefficients (for prediction problems) in the presence of data aggregation. Here, we identify a form of such bias in the estimation of R_0 that arises from the offset in temporal onset of epidemics in sub-regions used to produce aggregated epidemic data. Without fine-grained mobility information between sub-regions as in [52], the offset captures some degree of spatial information about how the virus spreads – for example, given sub-regional data, sorting the regions by their values of τ may prove useful in understanding how a new disease is spreading due to mobility. Closely related to our work in spirit is that of [53], who studies the effect of the incubation period for soilborne plant pathogens and the resulting differences in the understanding of the spread of plant diseases. Similarly, albeit with a different focus, [54] explore methods to model epidemic waves composed of overlapping sub-epidemics. They use generalized-logistic growth models to forecast trajectories of emerging epidemics. Among statistical approaches, [51] examined change in regression coefficients used for prediction as a function of data aggregation. Note also that Bayesian hierarchical models [55] have been used to capture knowledge about how epidemic curves in previous seasons might dictate the behavior of the epidemic in the following season. However, while [55] does experiment with a scaled and shifted SIR model, their analysis does not touch upon when and why such a model might be warranted, and the kinds of biases it corrects for.

Our work focuses on elucidating the effect of shift or delay between sub-units of data on the predictions done on the aggregated data. We showed how a mismatch between the generative assumptions of epidemiological models and how data is gathered can bias inferences made about the reproduction number and thus, the estimation of severity of epidemics. In particular, we quantified errors accrued when estimating R_0 in the presence of delayed epidemic onset as well as when the sub-regional epidemic data have delayed onsets. Our work illustrates how the lack of alignment between the assumptions made in a model and the data generating process can distort decisions made using the estimated parameters. To address and correct for this distortion, we introduced and validated the Shifted-SIR (S-SIR) model and provided an algorithm for parameter estimation, as a means to correct for the effect of delays in epidemic onset.

The subsequent analysis on synthetic data showcases the strength of the S-SIR model where we see that by explicitly accounting for time-delays within the compartmental model, we are able to correct and mitigate some of the bias in the estimation of R_0.

In its current form however the S-SIR model is limited to modeling time-shifted dynamics arising from an SIR model. Extending the approach to more general classes of compartmental models (e.g. the Susceptible-Exposed-Infected-Recovered model) may be appropriate for infectious disease that are known to undergo more complex dynamics.

When should the S-SIR model be used? If one has access to sub-regional local data, then policy decisions should be made separately for each sub-region. However, sub-regions can be sparsely populated, their data can be noisy or otherwise unavailable. In such scenarios policy decisions must be made from aggregated regional data. Recall from our earlier example, that serious heterogeneity in onset can introduce a systematic *under-estimation* of R_0. An under-estimation of 1.8 instead of 2.3 (Fig. 2) translates to a difference of 200,000 cases in a population of 1 million, in other words 20% error, which can be dramatic when estimating bed-capacity match to cases in time of pandemics. It is thus clear, that the best course of action would be to account for offsets with the S-SIR model when learning from regional – aggregated – epidemic curves. Doing so can mitigate some of the bias incurred in the estimation of R_0 with important implications for planning of response and management of cases.

Finally, as different modalities of data collection [56] are integrated for epidemiological modeling, it is vital to rethink our modeling framework to take into account the different levels of the data generation hierarchy [57]. Here, we presented one approach to mitigate errors, and systematic under-estimation in particular, in inferring R_0 due to heterogeneity in epidemic onset, particularly at early stages of epidemics when data is scarce and pooling data is common, while understanding of the underlying dynamics of the disease is not developed. Using our validated proposed S-SIR model and methodology, we hope that more accurate estimates of the reproduction number that matters on the ground can be used to improve planning and intervention, and mitigate mortality and morbidity rates in seasonal diseases such as influenza [58] and particularly at the early stages of newly emerging diseases such as those we are experiencing with COVID19 [6].

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Rahul G. Krishnan reports financial support was provided by Microsoft Research. Simone Cenci reports financial support was provided by DCI, LLC. This author has no additional relationships to disclose. The authors have no patents to disclose. The corresponding author reports no additional activities to disclose.

Acknowledgment

We acknowledge the support of the MIT Alumni Class Fund in support of HST.537/2.250/1.631 Fluids and Diseases course that enabled onset of this research, in addition to the support of the Burroughs Wellcome Fund, the Richard and Susan Smith Family Foundation, and the Wellcome Trust.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.annepidem.2021.07.008

References

[1] Bhutta ZA, Sommerfeld J, Lassi ZS, Salam RA, Das JK. Global burden, distribution, and interventions for infectious diseases of poverty. Infect Dis Poverty 2014;3(1):21.
[2] Blanton L, Dugan VG, Elal AIA, Alabi N, Barnes J, Brammer L, et al. Update: influenza activity-united states, September 30, 2018–february 2, 2019. Morbidity and Mortality Weekly Report 2019;68(6):125.
[3] Nair H, Brooks WA, Katz M, Roca A, Berkley JA, Madhi SA, et al. Global burden of respiratory infections due to seasonal influenza in young children: a systematic review and meta-analysis. The Lancet 2011;378(9807):1917–30.
[4] Centers for Disease Control (CDC). The burden of flu disease 2017–2018 infographic. https://www.cdc.gov/flu/resource-center/freeresources/graphics/flu-burden.htm,[Online; accessed 1-Nov-2019]; 2018a.
[5] Molinari N-AM, Ortega-Sanchez IR, Messonnier ML, Thompson WW, Wortley PM, Weintraub E, et al. The annual impact of seasonal influenza in the us: measuring disease burden and costs. Vaccine 2007;25(27):5086–96.
[6] Johns Hopkins University. Covid-19 dashboard by the center for systems science and engineering (csse) at johns hopkins university (jhu). https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6/; 2020.
[7] Jackson JK, Weiss MA, Schwarzenberg AB, Nelson RM. Global economic effects of Covid-19. Congressional Research Service; 2020.
[8] Cutler DM, Summers LH. The Covid-19 pandemic and the $16 trillion virus. JAMA 2020;324(15):1495.
[9] Mollison D. The structure of epidemic models. Epidemic models: their structure and relation to data 1995;1:17–33.
[10] Diekmann O, Heesterbeek JAP. Mathematical epidemiology of infectious diseases: model building, analysis and interpretation, 5. John Wiley & Sons; 2000.
[11] Huppert A, Katriel G. Mathematical modelling and prediction in infectious disease epidemiology. Clinical Microbiology and Infection 2013;19(11):999–1005.
[12] Metcalf CJE, Edmunds W, Lessler J. Six challenges in modelling for public health policy. Epidemics 2015;10:93–6.
[13] Choi B, Pak A. A simple approximate mathematical model to predict the number of severe acute respiratory syndrome cases and deaths. Journal of Epidemiology & Community Health 2003;57(10):831–5.
[14] Ferguson NM, Cummings DA, Cauchemez S, Fraser C, Riley S, Meeyai A, et al. Strategies for containing an emerging influenza pandemic in southeast asia. Nature 2005;437(7056):209.
[15] Hagmann R, Charlwood JD, Gil V, Ferreira C, Do Rosário V, Smith TA. Malaria and its possible control on the island of príncipe. Malar J 2003;2:15.
[16] Dietz K, Heesterbeek J. Daniel Bernoulli's epidemiological model revisited. Math Biosci 2002;180(1):1–21. doi:10.1016/S0025-5564(02)00122-0.
[17] Heffernan J, Smith R, Wahl L. Perspectives on the basic reproductive ratio. Journal of The Royal Society Interface 2005;2(4):281–93. http://rsif.royalsocietypublishing.org/content/2/4/281
[18] Brauer F. Mathematical epidemiology: past, present, and future. Infectious Disease Modelling 2017;2(2):113–27.
[19] Rebuli NP, Bean N, Ross J. Estimating the basic reproductive number during the early stages of an emerging epidemic. Theor Popul Biol 2018;119:26–36.
[20] Institute of Medicine. Ethical and legal considerations in mitigating pandemic disease: workshop summary. The National Academies Press; 2007. ISBN 978-0-309-10769-3. https://www.nap.edu/catalog/11917/ethical-and-legal-considerations-in-mitigating-pandemic-disease-workshop-summary
[21] Chowell G, Nishiura H, Bettencourt LM. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. Journal of The Royal Society Interface 2007;4(12):155–66. doi:10.1098/rsif.2006.0161. http://rsif.royalsocietypublishing.org/content/4/12/155.
[22] Fraser C, Donnelly CA, Cauchemez S, Hanage WP, Van Kerkhove MD, Hollingsworth TD, et al. Pandemic potential of a strain of influenza a (H1N1): early findings. Science 2009;324(5934):1557–61. doi:10.1126/science.1176062. http://science.sciencemag.org/content/324/5934/1557.
[23] Owada K, Eckmanns T, Kamara K-BO, Olu OO. Epidemiological data management during an outbreak of ebola virus disease: key issues and observations from sierra leone. Front Public Health 2016;4:163.
[24] Centers for Disease Control (CDC). Overview of influenza surveillance in the united states. https://www.cdc.gov/flu/weekly/overview.htm,[Online; accessed 1-Nov-2019]; 2017.
[25] Centers for Disease Control (CDC). 2018–2019 influenza season week 19 ending may 11, 2019. https://www.cdc.gov/flu/weekly/weeklyarchives2018-2019/Week19.htm,[Online; accessed 1-Nov-2019]; 2018b.
[26] Coltart CE, Lindsey B, Ghinai I, Johnson AM, Heymann DL. The ebola outbreak, 2013–2016: old lessons for new epidemics. Philosophical Transactions of the Royal Society B: Biological Sciences 2017;372(1721):20160297.
[27] Alter MJ, Mares A, Hadler SC, Maynard JE. The effect of underreporting on the apparent incidence and epidemiology of acute vhial hepatitis. Am J Epidemiol 1987;125(1):133–9.
[28] Dalziel BD, Lau MS, Tiffany A, McClelland A, Zelner J, Bliss JR, et al. Unreported cases in the 2014–2016 ebola epidemic: spatiotemporal variation, and implications for estimating transmission. PLoS Negl Trop Dis 2018;12(1):e0006161.
[29] Farrington CP, Kanaan MN, Gay NJ. Estimation of the basic reproduction number for infectious diseases from age-stratified serological survey data. Journal of the Royal Statistical Society: Series C (Applied Statistics) 2001;50(3):251–92.
[30] Schuessler AA. Ecological inference. Proceedings of the National Academy of Sciences 1999;96(19):10578–81.
[31] Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. Proceedings of the Royal Society A 1927;115(772):700–21. https://royalsocietypublishing.org/doi/10.1098/rspa.1927.0118.
[32] Larson RC. Simple models of influenza progression within a heterogeneous population. Oper Res 2007;55(3):399–412.
[33] Bourouiba L. Turbulent gas clouds and respiratory pathogen emissions: potential implications for reducing transmission of COVID-19. JAMA 2020;323:1837–8.
[34] Bourouiba L. Fluid dynamics of respiratory infectious diseases. Annu Rev Biomed Eng 2021;23(1):547–77.
[35] Brauer F. Compartmental models in epidemiology. Mathematical epidemiology 2008;1:19–79.
[36] Bourouiba L. Understanding the transmission of H5N1. CAB Reviews: Perspectives in Agriculture, Veterinary Sciences, Nutrition and Natural Resources 2013;8. 017:1–9
[37] Bourouiba L, Wu J, Newman S, Takekawa J, Natdorj T, Batbayar N, et al. Spatial dynamics of bar-headed geese migration in the context of H5N1. Journal of the Royal Society Interface 2010;7:1627–39.
[38] Bourouiba L, Teslya SL, Wu J. Highly pathogenic avian influenza outbreak mitigated by seasonal low pathogenic strains: insights from dynamic modeling. J Theor Biol 2011;271:181–201.
[39] Bourouiba L, Gourley S, Liu R, Takekawa J, Wu J. Avian influenza spread and transmission dynamics. mathematical modeling of infectious diseases. chapter 7 in analyzing and modeling spatial and temporal dynamics of infectious diseases. John Wiley & Sons; 2014.
[40] Centers for Disease Control (CDC). National, regional, and state level outpatient illness and viral surveillance. https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html,[Online; accessed 1-Nov-2019]; 2019.
[41] Yaari R, Katriel G, Huppert A, Axelsen J, Stone L. Modelling seasonal influenza: the role of weather and punctuated antigenic drift. Journal of The Royal Society Interface 2013;10(84):20130298.
[42] Centers for Disease Control (CDC). National notifiable diseases surveillance system (nndss). https://wwwn.cdc.gov/nndss/data-collection.html,[Online; accessed 1-Nov-2019]; 2015.
[43] Wikimedia Commons. US federal regions. https://commons.wikimedia.org/w/index.php?title=File:USFederalRegions.svg&oldid=307628742, [Online; accessed 9-December-2019]; 2018.
[44] Ferguson NM, Laydon D, Nedjati-Gilani G, Imai N, Ainslie K, Baguelin M, et al. Impact of non-pharmaceutical interventions (npis) to reduce Covid-19 mortality and healthcare demand. Imperial College COVID-19 Response Team 2020.
[45] Sun C, Hsieh Y-H. Global analysis of an seir model with varying population size and vaccination. Appl Math Model 2010;34(10):2685–97.
[46] Heffernan J, Wahl L. Improving estimates of the basic reproductive ratio: using both the mean and the dispersal of transition times. Theor Popul Biol 2006;70(2):135–45.
[47] Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling 2017;2(3):379–98.
[48] Keeling MJ, Eames KT. Networks and epidemic models. Journal of the Royal Society Interface 2005;2(4):295–307.
[49] Gou W, Jin Z. How heterogeneous susceptibility and recovery rates affect the spread of epidemics on networks. Infectious Disease Modelling 2017;2(3):353–67.

[50] Gehlke CE, Biehl K. Certain effects of grouping upon the size of the correlation coefficient in census tract material. J Am Stat Assoc 1934;29(185A):169–70.
[51] Clark WA, Avery KL. The effects of data aggregation in statistical analysis. Geogr Anal 1976;8(4):428–38.
[52] Sattenspiel L, Dietz K, et al. A structured epidemic model incorporating geographic mobility among regions. Math Biosci 1995;128(1):71–92.
[53] Leclerc M, Doré T, Gilligan CA, Lucas P, Filipe JA. Estimating the delay between host infection and disease (incubation period) and assessing its significance to the epidemiology of plant diseases. PLoS ONE 2014.
[54] Chowell G, Tariq A, Hyman JM. A novel sub-epidemic modeling framework for short-term forecasting epidemic waves. BMC Med 2019;17(1):1–18.
[55] Brooks LC, Farrow DC, Hyun S, Tibshirani RJ, Rosenfeld R. Flexible modeling of epidemics with an empirical bayes framework. PLoS Comput Biol 2015;11(8).
[56] Charles-Smith LE, Reynolds TL, Cameron MA, Conway M, Lau EH, Olsen JM, et al. Using social media for actionable disease surveillance and outbreak management: a systematic literature review. PLoS ONE 2015;10(10):e0139701.
[57] Mooney SJ, Westreich DJ, El-Sayed AM. Epidemiology in the era of big data. Epidemiology (Cambridge, Mass) 2015;26(3):390.
[58] Thompson WW, Weintraub E, Dhankhar P, Cheng P-Y, Brammer L, Meltzer MI, et al. Estimates of us influenza-associated deaths made using four different methods. Influenza Other Respir Viruses 2009;3(1):37–49.
[59] Berkson J, et al. Estimation by least squares and by maximum likelihood. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Contributions to the Theory of Statistics, 1; 1956.
[60] Branch MA, Coleman TF, Li Y. A subspace, interior, and conjugate gradient method for large-scale bound-constrained minimization problems. SIAM Journal on Scientific Computing 1999;21(1):1–23.
[61] Jones E., Oliphant T., Peterson P., et al. SciPy: Open source scientific tools for Python. 2001. http://www.scipy.org/.