Biostatistics, 2025, **26**(1), kxae054
https://doi.org/10.1093/biostatistics/kxae054
Article

Unlocking the power of time-since-infection models: data augmentation for improved instantaneous reproduction number estimation

Jiasheng Shi[1], Yizhao Zhou[2], Jing Huang [3,*]

[1]School of Data Science, , The Chinese University of Hong Kong, Shenzhen, 2001 Longxiang Boulevard, Shenzhen, 518172, China
[2]AstraZeneca, 1728-1746 West Nanjing Road, Shanghai, 200040, China
[3]Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, 423 Guardian Drive, Philadelphia, Pennsylvania, 19014, United States

[*]Corresponding author: Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, 625 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104, United States. Email: jing14@pennmedicine.upenn.edu

ABSTRACT

The time-since-infection (TSI) models, which use disease surveillance data to model infectious diseases, have become increasingly popular due to their flexibility and capacity to address complex disease control questions. However, a notable limitation of TSI models is their primary reliance on incidence data. Even when hospitalization data are available, existing TSI models have not been crafted to improve the estimation of disease transmission or to estimate hospitalization-related parameters—metrics crucial for understanding a pandemic and planning hospital resources. Moreover, their dependence on reported infection data makes them vulnerable to variations in data quality. In this study, we advance TSI models by integrating hospitalization data, marking a significant step forward in modeling with TSI models. We introduce hospitalization propensity parameters to jointly model incidence and hospitalization data. We use a composite likelihood function to accommodate complex data structure and a Monte Carlo expectation–maximization algorithm to estimate model parameters. We analyze COVID-19 data to estimate disease transmission, assess risk factor impacts, and calculate hospitalization propensity. Our model improves the accuracy of estimating the instantaneous reproduction number in TSI models, particularly when hospitalization data is of higher quality than incidence data. It enables the estimation of key infectious disease parameters without relying on contact tracing data and provides a foundation for integrating TSI models with other infectious disease models.

KEYWORDS: composite likelihood function; hospitalization propensity; infectious disease transmission; time between diagnosis and hospitalization.

1. INTRODUCTION

Mathematical modeling is essential for understanding infectious disease transmission, especially during the early stages of pandemics when vaccines, treatments, and immunity are limited. Policymakers rely on key metrics like the instantaneous reproduction number (R_t), projected incidence, and hospitalization counts to make timely decisions. Various modeling approaches, including

Received: February 8, 2024. **Revised:** December 17, 2024. **Accepted:** December 19, 2024

compartmental models, agent-based simulations, and time-series analyses, help estimate these metrics. Among these, time-since-infection (TSI) models have gained prominence due to their ability to estimate R_t directly from incidence data using renewal equation concepts.

The concept of TSI models originated in the 1910s by Ross (1916) and Ross and Hudson (1917a, b) and was mathematically formalized by Kermack and Ogilvy McKendrick (1927). They gained further prominence with contributions from Wallinga and Teunis (2004) and Fraser (2007). Recognition in the statistics community grew after the work of Cori et al. (2013) and the development of the EpiEstim R package (Cori 2021), which enabled widespread use for real-time R_t estimation, particularly during the COVID-19 pandemic (Gostic et al. 2020; Pan et al. 2020; Nash et al. 2022). Recent enhancements include regression integrations, addressing delays, and multi-location analyses (Quick et al. 2021; Shi et al. 2022; Ge et al. 2023). These models are based on the idea that new infections depend on three factors: recent infection counts, the current reproduction number, and how transmission evolves over time. They are mathematically linked to renewal equations, which model new cases as a sum of past infections weighted by infectious potential, and have similarities with branching processes that describe how each case leads to secondary infections (Athreya et al. 2004; Lloyd-Smith et al. 2005). Tools like the EpiNow2 package (Abbott et al. 2021) have applied these concepts to estimate time-varying transmission dynamics, accommodating various infectiousness patterns, and reporting delays.

Compared to other infectious disease models, e.g. compartmental models, TSI models are valued for their practicality and flexibility. Their statistical foundation enables integration with advanced methods, making them well-suited for complex modeling challenges. By relying primarily on empirical data, TSI models require minimal epidemiological knowledge, making them particularly useful during new infectious disease outbreaks with limited biological insights (Jewell 2021). A key advantage is their ability to estimate R_t directly from incidence data alone, making them user-friendly during early pandemic stages when other surveillance data are scarce. However, this reliance on incidence data can introduce biases from underreporting or reporting delays. Moreover, when other crucial data, such as hospitalization counts, become available, traditional TSI models are not equipped to integrate this information. It is because they assume each new infection is linked to prior cases, a pattern not followed by hospitalization data. This limitation restricts TSI models from effectively modeling disease-related hospitalizations, an essential metric for assessing pandemic trajectory and planning healthcare resources.

This study aims to address these limitations by developing an enhanced TSI model that integrates hospitalization data. We investigate whether incorporating hospitalization data can improve the precision of R_t estimates and offer insights into disease dynamics and severity in TSI models. Specifically, our approach aims to: (i) reduce bias in R_t estimation from incidence data inaccuracies, (ii) estimate hospitalization-related parameters for assessing disease severity, and (iii) combine the strengths of TSI and compartmental models for pandemic response. To achieve this, we introduce a new set of parameters, the hospitalization propensity, which quantifies the tendency of infectious individuals being hospitalized over TSI. By incorporating hospitalization data into the TSI framework, we use a composite likelihood function to jointly model incidence and hospitalization counts, capturing their interdependent dynamics (Lindsay 1988; Varin et al. 2011). The composite likelihood approach simplifies the modeling of complex data dependencies, allowing us to retain key statistical properties while reducing computational challenges. Using a Monte Carlo expectation–maximization (MCEM) algorithm, which combines stochastic sampling with iterative parameter estimation to handle missing data, we estimate model parameters and assess their accuracy through simulations and an analysis of COVID-19 data across U.S. counties. Our results demonstrate the enhanced precision of R_t estimation and highlight the capability of the proposed framework in supporting hospital resource planning and evaluating local transmission risks.

This work contributes to the field by unlocking the potential of TSI models through data augmentation of incidence and hospitalization data, which refines real-time estimation of R_t, enables the estimation of hospitalization-related parameters previously accessible primarily through contact tracing, and facilitates analysis of associations between risk factors and disease transmission

dynamics. The sections that follow outline the methodology, implementation strategy, and broader implications for infectious disease modeling.

2. METHODS

We first provide a brief overview of the fundamentals of TSI models, and then describe the proposed approach to integrate incidence data with hospitalization data based on the TSI framework.

2.1. Model and notations

We use subscripts t to denote calendar time and s to indicate time since infection. Consider a discrete-time setting where I_t represents the number of new infections on day t. Given the infection history up to day $t-1$, the expected new infections on day t can be expressed as $E(I_t|I_0,\ldots,I_{t-1}) = R_t\Lambda_t$, where $\Lambda_t = \sum_{s=1}^{t} I_{t-s}\omega_s$ represents the infection potential on day t. This potential is shaped by the current number of infectious individuals and the infectiousness function ω_s, which describes how infectious an individual is s days after infection, with $\sum_{s=0}^{+\infty}\omega_s = 1$. Typically, ω_s is set to zero for $s = 0$ and for $s > \eta$, where η is the time between infection and recovery. This function is often approximated using the distribution of the serial interval or generation time (Svensson 2007), and estimating R_t involves assuming a distribution for I_t and applying methods like maximum likelihood or Bayesian approaches with predefined values of ω_s (Cori et al. 2013; WHO Aylward et al. 2014; Quick et al. 2021).

By "hospitalization," we refer to hospital admission due to the infectious disease. Let H_t denote the number of new hospitalizations on day t. We define the filtrations $\mathcal{F}_t = \sigma(\{I_r, 0 \le r \le t\})$ and $\mathcal{G}_t = \sigma(\{I_r, H_r, 0 \le r \le t\})$ which represent the information on past infections and the combined information on both past infections and hospitalizations, respectively. In line with the original TSI models, we assume the number of new infections at time t, given the number of previously infected individuals, follows a Poisson distribution:

$$I_t|\mathcal{F}_{t-1} \sim \text{Poisson}(R_t\Lambda_t), \tag{2.1}$$

Next, we introduce a new set of parameters, the hospitalization propensity of an infected individual, $\tilde{\omega}_s$, which describes the tendency that an infected individual will be hospitalized s days after infection, independent of the calendar time t. For each infected individual, $\tilde{\omega}_s$ represents the probability that hospitalization occurs s days after infection. For the entire infected population, it represents the proportion of individuals hospitalized s days post-infection. Similar to $\{\omega_s\}$, the set $\{\tilde{\omega}_s\}$ satisfies the conditions $\tilde{\omega}_s \ge 0$ and $\sum_s \tilde{\omega}_s = 1$.

We denote $h_{t,s}$ as the number of patients infected at calendar time t and admitted at calendar time $t+s$, i.e. time s since infection, with $\mathbb{E}(h_{t,s}|\mathcal{F}_t) = \tilde{\omega}_s I_t$. If hospitalization does not occur within a certain period $\tilde{\eta}$, we assume $h_{t,s} = 0$ and $\tilde{\omega}_s = 0$ for $s > \tilde{\eta}$ for all t, where the positive integer $\tilde{\eta}$ represents the duration from infection to the time after which the likelihood of hospitalization becomes negligible. Patients infected at time t and never hospitalized are denoted by $h_{t,-1}$ with the tendency $\tilde{\omega}_{-1} = 1 - \sum_{s=0}^{\tilde{\eta}} \tilde{\omega}_s$. Thus, total incidence I_t and total hospitalizations H_t at time t are connected through

$$I_t = h_{t,-1} + \sum_{s\ge 0} h_{t,s}, \text{ and } H_t = \sum_{s=0}^{t} h_{t-s,s}, \text{ for } t = 0,1,2,\cdots. \tag{2.2}$$

Based on the above setup, we further assume

$$(h_{t,-1}, h_{t,0}, \cdots, h_{t,\tilde{\eta}})|I_t \sim \text{Multinomial}(I_t, \tilde{\omega}_{-1}, \tilde{\omega}_0, \cdots, \tilde{\omega}_{\tilde{\eta}}). \tag{2.3}$$

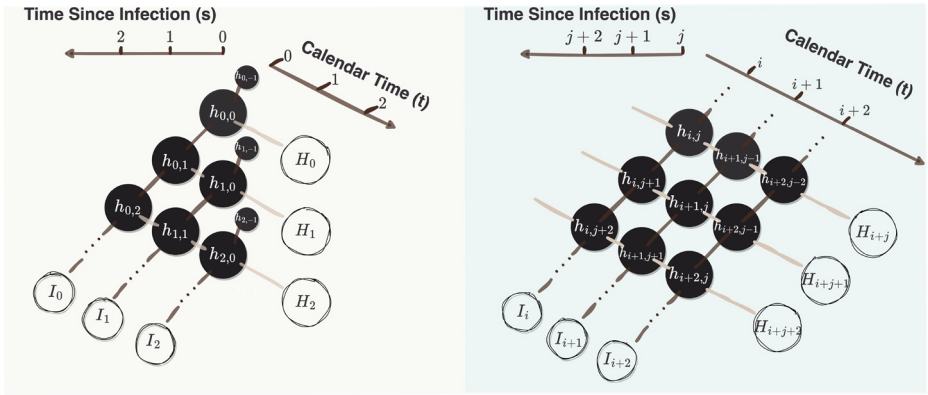

Fig. 1. The relationship between incidence data I_t and hospitalization data H_t. The left panel illustrates the relationship between calendar time 0 and 2 and TSI from 0 to 2. The right panel illustrates the relationship between calendar time i and $i + 2$ and TSI from j to $j + 2$.

From (2.2) and (2.3), we can see the time-series infection and hospitalization data are deeply interconnected and overlapping (Fig. 1). To model R_t in relation to risk factors, we assume a regression structure inspired by Shi et al. (2022) and Quick et al. (2021):

$$f_{link}(R_t) = X_t^T \beta + \sum_{i=1}^{q} \theta_i f_i(D_{1,t}, D_{2,t}, D_{3,t}), \tag{2.4}$$

where $t \geq q > 0$ and X_t is a p-dimensional vector that includes a constant 1 and a vector of risk factors of disease transmission, Z_t, such as temperature, social distancing measures, and population density. Here, β represents the effect size of these exogenous variables on transmission. The link function $f_{link}(\cdot)$ transforms R_t, allowing it to be expressed as a linear combination of predictors, e.g. the log link function is one of the most commonly used link functions for non-negative outcomes, while $f_i(\cdot)$ represent a functions of past outcomes. Additionally, we define

$$D_{1,t} = \{X_r\}_{0\leq r\leq t}, \quad D_{2,t} = \{I_r, \mathbb{E}(H_r)\}_{0\leq r\leq t-1}, \quad D_{3,t} = \{R_r\}_{0\leq r\leq t-1}. \tag{2.5}$$

With $D_{1,t}$ being observed, (2.4) ensures that $R_t \in \mathcal{F}_t$. Therefore, the second term in (2.4) captures the temporal structure of disease transmission, while the θ_i's characterize the level of temporal dependency and continuity in the time series of $\{R_t\}_{t\geq 0}$. By combining (2.1) to (2.5), we build a TSI model for hospitalization and incidence data. Indeed, the terms $\{\mathbb{E}(H_r)\}_{1\leq r\leq t-1}$ in $D_{2,t}$ can be expressed as functions of the hospitalization propensity and the R_t according to our model. This implies $D_{2,t} \subset \sigma(\{I_r, R_r\}_{0\leq r\leq t-1})$ and hospitalization data are linked with the disease transmission through the hospitalization propensity. In a broader context, one could substitute$D_{2,t}$ in (2.5) with $\{I_r, H_r\}_{0\leq r\leq t-1}$. We have reserved this general setting to the Supplementary Materials, focusing here on the model shown in (2.5).

2.2. Composite likelihood function

Specifying the full likelihood function for the proposed model is challenging. Specifically, the joint distribution of (I_r, H_r) given $\mathcal{G}_{r-1}$ for $0 \leq r \leq t$ is difficult to retrieve due to the convolution structure outlined in (2.2), the irregular boundaries and range restrictions applied to $\{I_r, H_r\}_{0\leq r\leq t}$, and the fact that $h_{t,s}$ are often unobserved in practice. Therefore, we adopt a composite likelihood

approach to estimate the model parameters. Specifically, we use joint distribution of (I_r, H_r) given $\mathcal{F}_{r-1}$ instead of $\mathcal{G}_{r-1}$ to construct a composite log-likelihood function

$$\ell_C = \sum_{0\le r\le t} \log \mathbb{P}(H_r, I_r|\mathcal{F}_{r-1}). \tag{2.6}$$

We derive (2.6) by summing over all possible values of $\{h_{r-s,s}\}_{1\le s\le \tilde{\eta}}$ in the joint distribution of (H_r, I_r) and $\{h_{r-s,s}\}_{1\le s\le \tilde{\eta}}$ conditioning on $\mathcal{F}_{r-1}$. To achieve this, we first demonstrate in the following lemma that $h_{t,s}$ given $\mathcal{F}_{t-1}$ follows a Poisson distribution and that h_{t,u_1} and h_{t,u_2} are independent given $\mathcal{F}_{t-1}$ when $u_1 \neq u_2$. Then we derive the form of individual component $\mathbb{P}(H_r, I_r|\mathcal{F}_{r-1})$ based on this lemma.

Lemma 1 *For arbitrary* $t > 0$, $-1 \le u_1, u_2 \le \min\{t, \tilde{\eta}\}$ *and* $u_1 \neq u_2$, *we have*

$$h_{t,u_1}|\mathcal{F}_{t-1} \sim Poisson(\tilde{\omega}_{u_1} R_t \Lambda_t), \quad h_{t,u_2}|\mathcal{F}_{t-1} \sim Poisson(\tilde{\omega}_{u_2} R_t \Lambda_t), \tag{2.7}$$

and $h_{t,u_1}|\mathcal{F}_{t-1} \perp h_{t,u_2}|\mathcal{F}_{t-1}$. *Moreover,* $(h_{t,u_1}, h_{t,u_2})|\mathcal{G}_{t-1} = (h_{t,u_1}, h_{t,u_2})|\mathcal{F}_{t-1}$, *and* $h_{t,u_1}|\mathcal{G}_{t-1} \perp h_{t,u_2}|\mathcal{G}_{t-1}$.

According to (2.3) and Lemma 1, if we use $\mathbf{1}(\cdot)$ to denote the indicator function, which equals 1 when its condition is true and 0 otherwise, the joint distribution of (H_r, I_r) and $\{h_{r-s,s}\}_{1\le s\le \tilde{\eta}}$ conditioning on $\mathcal{F}_{r-1}$ can be expressed as the product of the probability density functions of several binomial distributions and two Poisson distributions as follows:

$$\begin{aligned}
&\mathbb{P}(H_r, I_r, h_{r-\tilde{\eta},\tilde{\eta}}, \cdots, h_{r-1,1}|\mathcal{F}_{r-1}) \\
&= \mathbb{P}(h_{r-\tilde{\eta},\tilde{\eta}}, \cdots, h_{r-1,1}, h_{r,0} = H_r - \sum_{s=1}^{\tilde{\eta}} h_{r-s,s}, \sum_{s=-1,1,\cdots,\tilde{\eta}} h_{r,s} = I_r - H_r + \sum_{s=1}^{\tilde{\eta}} h_{r-s,s}|\mathcal{F}_{r-1}) \\
&= \prod_{s=1}^{\tilde{\eta}} \mathbb{P}(Binomial(I_{r-s}, \tilde{\omega}_s) = h_{r-s,s}) \mathbf{1}(H_r - I_r \le \sum_{s=1}^{\tilde{\eta}} h_{r-s,s} \le H_r) \\
&\quad \cdot \mathbb{P}(Poisson(\tilde{\omega}_0 R_r \Lambda_r) = H_r - \sum_{s=1}^{\tilde{\eta}} h_{r-s,s}) \\
&\quad \cdot \mathbb{P}(Poisson((1-\tilde{\omega}_0) R_r \Lambda_r) = I_r - H_r + \sum_{s=1}^{\tilde{\eta}} h_{r-s,s})
\end{aligned} \tag{2.8}$$

Therefore, the joint distribution of (H_r, I_r) conditioning on $\mathcal{F}_{r-1}$ can be calculated by summing over all possible values of $\{h_{r-s,s}\}_{1\le s\le \tilde{\eta}}$,

$$\mathbb{P}(H_r, I_r|\mathcal{F}_{r-1}) = \sum_{h_{r-\tilde{\eta},\tilde{\eta}}, \cdots, h_{r-1,1}} \mathbb{P}(H_r, I_r, h_{r-\tilde{\eta},\tilde{\eta}}, \cdots, h_{r-1,1}|\mathcal{F}_{r-1}). \tag{2.9}$$

3. INFERENCE

3.1. Estimation

In this section, we describe the procedures for estimating model parameters using MCEM algorithms, where the missing data $\{h_{r-s,s}\}_{1\le s\le \tilde{\eta}}$ are unobserved tracing information and the observed disease data consist of daily incidence and hospitalization counts. Let $\gamma = (\beta, \theta, \omega, \tilde{\omega}) \in \Gamma$ denote the model parameters, where β and θ define a time series model for the instantaneous reproduction

number $\{R_t\}_{t\geq 1}$, as described in (2.4). The terms $\{R_t\}_{t\geq 1}$ are integrated into the composite likelihood function (2.6) through (2.8) and (2.9), introducing β and θ as parameters to be estimated. We use notations like $\mathbb{P}_\gamma$ to indicate that the probability is associated with the parameter value γ. We use γ_0 to denote the parameter values corresponding to the underlying true data-generating mechanism, and $\hat{\gamma}$ denote the maximum composite likelihood estimator. The composite likelihood function integrates information from both the observed data and the modeled structure of $\{R_t\}_{t\geq 1}$, and the EM algorithm addresses the missing data, enabling the joint estimation of all parameters γ. To improve the efficiency of sampling the missing data $h_{r-\tilde{\eta},\tilde{\eta}}, \cdots, h_{r-1,1}$ in the algorithm, we incorporate an acceptance-rejection sampling method.

Let $Data_{obs,r} = \{X_j, I_j\}_{0\leq j\leq r} \cup \{H_r\}$ denote the observed data and $Data_{miss,r} = \{h_{r-s,s}\}_{1\leq s\leq \tilde{\eta}}$ denote the missing data at time r. In the MCEM algorithm, N_0 denotes the Monte-Carlo sample size, and $Data_{miss,r}^{(m,k)}$, $1 \leq m \leq N_0$ denotes the m-th Monte-Carlo sample in the k-th iteration of the EM algorithm. Throughout the following, (k) denotes the k-th EM iteration. For example, $\gamma^{(k)}$ represents the current estimate of γ after the k-th EM iteration, with $R_r^{(k)}$ and $\Lambda_r^{(k)}$ representing estimates of R_r and Λ_r based on (2.1) and (2.4), using parameter values $\gamma^{(k)}$. The estimation procedure is outlined below, with a pseudo-code provided in Algorithm 1:

E step At the $(k+1)$-th iteration, given the current estimate $\gamma^{(k)} = (\beta^{(k)}, \theta^{(k)}, \omega^{(k)}, \tilde{\omega}^{(k)})$, this step computes the expected complete data composite log-likelihood, or the Q-function:

$$Q(\gamma|\gamma^{(k)}) \stackrel{\text{def}}{=} \sum_{0\leq r\leq t} \mathbb{E}_{\gamma^{(k)}}(\log \mathbb{P}_\gamma(H_r, I_r, Data_{miss,r}|\mathcal{F}_{r-1})|Data_{obs,r}), \tag{3.10}$$

where the expectation is taken over the distribution of $Data_{miss,r}$ conditional on $(Data_{obs,r}, \gamma^{(k)})$. Due to computational complexity, the Q-function is approximated using Monte Carlo sampling of $Data_{miss,r}$ from $\mathbb{P}_{\gamma^{(k)}}(Data_{miss,r}|Data_{obs,r})$, and is calculated as

$$\hat{Q}(\gamma|\gamma^{(k)}) \stackrel{\text{def}}{=} \sum_{0\leq r\leq t} \frac{1}{N_0} \sum_{m=1}^{N_0} \log \mathbb{P}_\gamma(H_r, I_r, Data_{miss,r}^{(m,k)}|\mathcal{F}_{r-1}).$$

M step This step is to compute $\gamma^{(k+1)} = \arg\max_{\gamma\in\Gamma} \hat{Q}(\gamma|\gamma^{(k)})$. If $\arg\max_{\gamma\in\Gamma} \hat{Q}(\gamma|\gamma^{(k)})$ is not unique, we randomly choose one as $\gamma^{(k+1)}$. When $\hat{Q}(\gamma^{(k)}|\gamma^{(k)}) = \max_{\gamma\in\Gamma} \hat{Q}(\gamma|\gamma^{(k)})$, we choose $\gamma^{(k+1)} = \gamma^{(k)}$.

Notably, drawing samples from $P_{\gamma^{(k)}}\left(Data_{\text{miss},r}|Data_{\text{obs},r}\right)$ is equivalent to sampling from $P_{\gamma^{(k)}}\left(H_r, I_r, Data_{\text{miss},r}|\mathcal{F}_{r-1}\right)$, as $P_{\gamma^{(k)}}(H_r, I_r|\mathcal{F}_{r-1})$ in (2.9) is constant for a given $\gamma^{(k)}$. To improve sampling efficiency, we use an acceptance-rejection method in Algorithm 1. Specifically, samples are first drawn from $P_{\gamma^{(k)}}\left(Data_{\text{miss},r}|\mathcal{F}_{r-1}\right)$ and accepted with probability $p_{\text{acceptance}}$, proportional to $P_{\gamma^{(k)}}\left(H_r, I_r|Data_{\text{miss},r}, \mathcal{F}_{r-1}\right)$, where $\max_{H_r,I_r} p_{\text{acceptance}} \leq 1$. To further enhance efficiency, particularly when infections and hospitalizations are large $(H_t \geq 25)$ in later pandemic stages, $p_{\text{acceptance}}$ in Algorithm 1 can be adjusted using Stirling's approximation:

$$p_{\text{acceptance-adj}} = 2\pi R_r^{(k)} \Lambda_r^{(k)} \sqrt{\tilde{\omega}_0^{(k)}(1-\tilde{\omega}_0^{(k)})} p_{\text{acceptance}}.$$

In Algorithm 1, the parameters of the infectiousness function, ω_s, and hospitalization propensity, $\tilde{\omega}_s$, are assumed to be unknown. This scenario is often encountered in the early stages of novel pandemics when little is known about the infectious pathogens. However, prior knowledge about ω_s and $\tilde{\omega}_s$ may sometimes be available from multiple biomedical or contact tracing studies. In

Algorithm 1 An MCEM algorithm to estimate model parameters with unknown ω_s and $\tilde{\omega}_s$.

Require: initial parameter value $\gamma^{(0)}$, $k \leftarrow 0$, breakpoint critical value Δ_0.
  while $||\gamma^{(k+1)} - \gamma^{(k)}||_\infty > \Delta_0$, **set** $r \leftarrow 0$, $m \leftarrow 1$, **do**
    while $0 \le r \le t$, $1 \le m \le N_0$, **do**
      Sample $h_{r-s,s}$ independently from Binomial$(I_{r-s}, \tilde{\omega}_s^{(k)})$, for $1 \le s \le \tilde{\eta}$.
      Sample ψ from Bernoulli distribution with probability $p_{\text{acceptance}}$, where
      $p_{\text{acceptance}} = \mathbb{P}(Poisson(\tilde{\omega}_0^{(k)} R_r^{(k)} \Lambda_r^{(k)}) = H_r - \sum_{s=1}^{\tilde{\eta}} h_{r-s,s})$
      $\times \mathbb{P}(Poisson((1-\tilde{\omega}_0^{(k)}) R_r^{(k)} \Lambda_r^{(k)}) = I_r - H_r + \sum_{s=1}^{\tilde{\eta}} h_{r-s,s})$
      if $\psi = 1$, **then**
        let $Data_{miss,r}^{(m,k)} = \{h_{r-s,s}\}_{1 \le s \le \tilde{\eta}}$, $m \leftarrow m + 1$.
      end if
      if $m > N_0$, **then**
        let $r \leftarrow r + 1$, $m \leftarrow 1$.
      end if
    end while
    Calculate the Monte Carlo Q-function $\hat{Q}(\gamma \,|\gamma^{(k)})$ and $\gamma^{(k+1)} = \arg\max_{\gamma \in \Gamma} \hat{Q}(\gamma \,|\gamma^{(k)})$.
    Let $k \leftarrow k + 1$, $\hat{\gamma} \leftarrow \gamma^{(k+1)}$.
  end while
  Output $\hat{\gamma}$.

such scenarios, we may consider these estimates from other studies as prior knowledge. In the Supplementary Materials, we provide another MCEM algorithm, Algorithm 2, to estimate the model parameters when prior knowledge about ω_s and $\tilde{\omega}_s$ is available.

3.2. Asymptotic properties

The basic TSI model (2.1), which describes the generative nature of an infectious pathogen, shares similarities with a branching process. Since branching process degenerates on an extinction set with no asymptotic properties, we similarly define an extinction set $\mathcal{E}$ for the TSI model,

$$\mathcal{E} \overset{\text{def}}{=} \Big\{ I_r = 0, \text{ for } r \text{ greater than some } K \Big\} = \bigcup_{r \ge 1} \Big\{ \frac{\partial \log \mathbb{P}(I_r, H_r \mid \mathcal{F}_{r-1})}{\partial \gamma} = 0 \Big\}, \tag{3.11}$$

and define $\mathcal{E}_{none} = \mathcal{E}^c$. Often, no consistent estimator $\hat{\gamma}$ exists in $\mathcal{E}$. Thus, we set aside the extinction probability

$$\mathbb{P}(\mathcal{E}) = \mathbb{P}\left(\bigcup_{r \ge 1} \Big\{ \frac{\partial \log \mathbb{P}(I_r, H_r | \mathcal{F}_{r-1})}{\partial \gamma} = 0 \Big\} \right)$$

and focus on the asymptotic behavior of $\hat{\gamma}$ on the non-extinction set $\mathcal{E}_{none}$.

In the following theorems, we show that the proposed MCEM algorithm preserves the ascent property of the composite likelihood function and converges for our model. We then establish the consistency of the maximum composite likelihood estimator and present it in the practical form of (2.4), under certain regularity conditions. Proof details are provided in the Supplementary Materials.

Theorem 1 (Ascent property of the composite likelihood) *For $k \ge 0$, we have $\ell_C(\gamma^{(k+1)}) \ge \ell_C(\gamma^{(k)})$.*

Theorem 2 (Convergence of the MCEM algorithm) *Assume Γ is a compact set, then with $k \to \infty$, the MCEM-estimator $\gamma^{(k)}$ at the k-th iteration converges to one of the stationary point γ_s induced by $M(\cdot) = \arg\max_{\gamma \in \Gamma} Q(\gamma \,|\cdot)$, and $\ell_C(\gamma^{(k)})$ converges monotonically to $\ell_C(\gamma_s)$.*

To establish the consistency of the maximum composite likelihood estimator, we impose regularity conditions on the observed time series. Inspired by work on counting processes (Zeger and Qaqish 1988; Davis et al. 1999, 2000, 2003), we assume:

Condition 1 (Series ergodicity) *There exist* $t_0 > 0$, *such that,*

$$\lim_{t\to\infty} \frac{t_0}{t} \sum_{s=1}^{t} R_s \Lambda_s \to_{a.s.} \sum_{s=1}^{t_0} \mathbb{E} R_s \Lambda_s, \quad \lim_{t\to\infty} \frac{t_0}{t} \sum_{s=1}^{t} I_s \log(R_s) \to_{a.s.} \sum_{s=1}^{t_0} \mathbb{E} I_s \log(R_s).$$

We also assume the time series regression in (2.4) uses a log link function and follows an autoregressive model of order 1, such that $\log(R_t) = Z_t^T \beta + \theta_0 + \theta_1 \log(R_{t-1})$, where θ_0 is an intercept term extracted from both the exogenous terms and the autoregressive terms and θ_1 serves as the autoregressive parameter, capturing the dependence of R_t on its previous value. We then establish the consistency of the maximum composite likelihood estimator for the time series regression coefficients when the infectiousness function and hospitalization propensity, $(\omega_s, \tilde{\omega}_s)$, are known. We denote the parameters of interest as $\gamma\,|_\omega = (\beta, \theta)$. The consistency of the maximum composite likelihood estimator, $\hat{\gamma}\,|_\omega = (\hat{\beta}, \hat{\theta})$, is shown as follows.

Theorem 3 *(Strong consistency of the estimators) Under condition 1 and assuming that* Γ *is a compact set, we have* $\hat{\gamma}\,|_\omega \xrightarrow{a.s.} \gamma_0|_\omega$ *as the length of the observational days* $t \to \infty$, *where* $\gamma_0|_\omega$ *is the parameter value corresponding to the true data generating mechanism.*

Furthermore, with the following conditions from the theorem 3 of Kaufmann (1987), i.e.

Condition 2 (Non-singularity) *For each* $1 \le r \le t$, *define*

$$\xi_r(\gamma\,|_\omega) = \frac{\partial R_r}{\partial \gamma\,|_\omega} \cdot R_r^{-1} \cdot (I_r - R_r \Lambda_r).$$

There exists some nonrandom and non-singular normalizing matrix A_t, *such that the normalized conditional variance converges to an almost surely positive definite random matrix* $\zeta^T \zeta$, *i.e.*

$$A_t^{-1} \left[\sum_{r=1}^{t} \mathrm{Cov}(\xi_r(\gamma_0|_\omega) | \mathcal{F}_{r-1}) \right] (A_t^{-1})^T \xrightarrow{P} \zeta^T \zeta.$$

Condition 3 (Uniformly integrability) *For* $1 \le r \le t$, $\mathbb{E}[\xi_r^2(\gamma_0)|\mathcal{F}_{r-1}]$ *is termwise uniformly integrable.*

Condition 4 (The conditional Lindeberg condition) *For arbitrary* $\epsilon > 0$,

$$\sum_{r=1}^{t} \mathbb{E}[\xi_r^T(\gamma_0|_\omega)(A_t^T A_t)^{-1} \xi_r(\gamma_0|_\omega) \cdot \mathbf{1}(|\xi_r^T(\gamma_0|_\omega)(A_t^T A_t)^{-1} \xi_r(\gamma_0|_\omega)| > \epsilon^2) | \mathcal{F}_{r-1}] \xrightarrow{P} 0.$$

Condition 5 (The smoothness condition) *For arbitrary* $\delta > 0$, *and* $\delta-$*neighborhood ball* $\mathcal{B}_t(\delta)$ *defined as* $\mathcal{B}_t(\delta) = \{\tilde{\gamma}\,|_\omega : ||A_t^T(\tilde{\gamma}\,|_\omega - \gamma_0|_\omega)|| \le \delta\}$,

$$\sup_{\tilde{\gamma}\,|_\omega \in \mathcal{B}_t(\delta)} ||A_t^{-1} \sum_{r=1}^{t} (\frac{\partial \xi_r(\tilde{\gamma}\,|_\omega)}{\partial \gamma} + \mathrm{Cov}(\xi_r(\gamma_0|_\omega)|\mathcal{F}_{r-1}))(A_t^{-1})^T|| \xrightarrow{P} 0.$$

the maximum composite likelihood estimator converges to a normal distribution as shown below.

Theorem 4 (Asymptotic normality) *Under condition 1-5, and assume* Γ *is a compact set, then on the non-extinction set defined in (3.11), with* $t \to \infty$,

$$[\sum_{r=1}^{t} \text{Cov}\,(\xi_r(\gamma_0|_\omega)|\mathcal{F}_{r-1})]^{1/2}(\hat{\gamma}\,|_\omega - \gamma_0|_\omega) \xrightarrow{d} N(0, I), \tag{3.12}$$

where I is the identity matrix.

4. SIMULATION STUDIES

4.1. Simulation setups

To evaluate the performance of the proposed method, we conducted simulation studies under two sets of scenarios. In the first, reported new infections represent the true counts, meaning the TSI model is correctly specified (correctly specified model). In the second, reported counts include errors (e.g. under-reporting or limited testing), implying the TSI model is misspecified (misspecified model). In both scenarios, hospitalization data were assumed accurate. Each scenario was repeated 1,000 times to assess bias and coverage probability of the estimator.

Daily infections, hospitalizations, and covariates were generated from the proposed model (2.1)-(2.5), with a simplified form of (2.4): $\log(R_r) = Z_r^T\beta + \theta_0 + \theta_1 \log(R_{r-1})$, for $r \geq 1$, which means that the logarithm of the instantaneous reproduction number $\{R_r\}_{1\leq r\leq t}$ follows an AR(1) structure with exogenous terms. Following Li et al. (2020), the infectiousness function, approximated by the serial interval, was reported to follow a Gamma distribution with a mean of 7.5 d (95% CI: [5.3, 19]), corresponding to a standard deviation ranging from approximately 1.12 to 5.86. Our simulations used a Gamma distribution for the infectiousness function, with shape and scale parameters adjusted to mirror these results. Similarly, hospitalization propensity was set to a Gamma distribution to mimic the delay between disease onset and hospitalization reported in the same study:

$$\omega_s = \mathbb{P}(\Gamma(k_1, \mu_1) \in [s-1, s)),\ \ s = 1, \cdots, 24, \quad \text{and}\ \ \omega_{25} = \mathbb{P}(\Gamma(k_1, \mu_1) \geq 24),$$
$$2\tilde{\omega}_s = \mathbb{P}(\Gamma(k_2, \mu_2) \in [s, s+1)),\ s = 0, \cdots, 4,\ 2\tilde{\omega}_5 = \mathbb{P}(\Gamma(k_2, \mu_2) \geq 5),\ \ \text{and}\ \tilde{\omega}_{-1} = 0.5,$$

where $\Gamma(\cdot,\cdot)$ had shape parameters $\check{k}_1 = 2.5$, $\check{k}_2 = 1.6$, and scale parameters $\mu_1 = 3$, $\mu_2 = 1.5$. We set the study duration $T = 120$ and fixed $p = 2$. Parameter values were assigned as $(\theta_0, \theta_1, \beta^T) = (0.7, 0.5, -.02, -.125)$. We independently simulated two =covariates $\{Z_{r,1}, Z_{r,2}\}_{1\leq r\leq t}$ to mimic real data on temperature in Philadelphia and social distancing trends obtained from daily cellular telephone movements, as provided by Unacast (Unacast, July 1st). This data represented the percentage change in visits to non-essential businesses, such as restaurants and hair salons, between March 1st and June 30th, 2020.

4.2. Simulation results with correctly specified model

In this set of scenarios, we explored three circumstances and compared the proposed method to the reference approach by Cori et al. (2013). In the first circumstance, we assumed the infectiousness function ω_s and hospitalization propensity $\tilde{\omega}_s$ were known. The model parameters reduced to $\gamma = (\theta^T, \beta^T)$. For the reference approach, a 3-d sliding window was chosen by minimizing the $\mathcal{L}_2$-distance between the estimated and oracle $\{R_r\}_{1\leq r\leq t}$. Additional results for 1-d and 7-d windows are in the Supplementary Materials. For the proposed method, we selected the initial $\gamma^{(0)}$ randomly and away from the oracle value. Figure 2 shows both methods captured the trend of $\{R_r\}_{1\leq r\leq t}$ well with very small estimation biases, but the proposed method performed better than the baseline method with a smaller bias. Moreover, when ω_s and $\tilde{\omega}_s$ were known, we found the proposed MCEM algorithm converged very fast, and the estimation bias for $\{R_r\}_{1\leq r\leq t}$ reached its limit after only two iterations of running the algorithm. In the second circumstance, we assumed

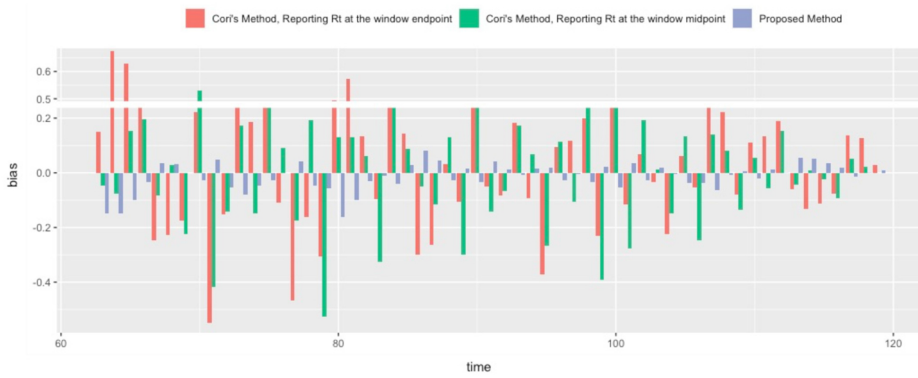

Fig. 2. Comparison of the estimation bias in the daily instantaneous reproduction number, R_t, between the proposed method, Cori's method that reports R_t at the endpoint of the sliding window (Cori et al. 2013), and Cori's method that reports R_t at the midpoint of the sliding window (Gostic et al. 2020; Gressani et al. 2022), when the reported number of daily new infections was accurate, and both the infectiousness function ω_s and hospitalization propensity $\tilde{\omega}_s$ were known. For the reference approach, a 3 d sliding window is selected by minimizing the $\mathcal{L}_2$-distance of the estimated and oracle sequence of $\{R_r\}_{1\leq r\leq t}$, and a comparison to reference approach with different sliding window is left to the supplementary. The estimation bias was calculated as the difference between the estimated daily R_t and the oracle R_t that generates the data. The bias is plotted starting from day 60, as Cori et al. (2013) demonstrated substantial bias and unstable estimation during the early stages of the simulated period when the incident cases I_t were small ($\leq 2 \times 10^3$).

ω_s and $\tilde{\omega}_s$ were unknown but prior knowledge about ω_s was available. We used the estimated infectiousness functions from previous studies as the prior knowledge (ALi et al. 2020; Wu et al. 2020; Deng et al. 2021; Chen et al. 2022), set the true function according to the results in Li et al. (2020), and used Algorithm 2 to estimate the parameters. Results were similar to those observed in the first circumstance, except for a small increase in estimation bias for both methods. In the third circumstance, we assumed ω_s and $\tilde{\omega}_s$ were unknown and no prior knowledge was available. In this situation, we only estimated the parameters using the proposed method, since the reference approach either requires user-specified values to constrain the overall shape of the infectiousness function or needs user-provided contact tracking data to estimate the infectiousness function. Using the proposed method, we estimated the regression coefficients as well as the infectiousness function and hospitalization propensity. As shown in Table 1, the proposed method produced accurate estimates for parameters with small bias and good coverage probability. It also consistently estimated the instantaneous reproduction numbers (Fig. S7). Overall, when the TSI model was correctly specified, the proposed composite likelihood MCEM algorithm benefited from incorporating hospital admission data, outperforming the reference approach.

4.3. Simulation results with misspecified model

In the second set of scenarios, we assumed the reported daily new infections were inaccurate due to underreporting, with the proportion of reported cases varying daily. This reflects real-world challenges during the pandemic, where case reporting was influenced by testing availability, public compliance, and at-home testing. In contrast, hospitalization data were assumed accurate, as hospitals report admissions in standardized formats. Additional simulations involving underreporting of hospitalizations are included in the Supplementary Materials.

Daily underreporting percentage was generated from a normal distribution with a mean of 15% and standard deviation of 5%. This allowed the daily under-reporting percentage to vary from 0% to 30%, resulting in poor data quality for the reported infection numbers. However, we assumed the

Table 1. Performance of the proposed method when the reported number of daily new infections was accurate, the infectiousness function ω_s and hospitalization propensity $\tilde{\omega}_s$ were unknown and no prior knowledge was available.[a]

	θ_0	θ_1	β_1	β_2	ω_5	ω_6	$\tilde{\omega}_0$
Empirical bias ($\times 10^{-3}$)	−0.15	2.10	0.08	0.15	−0.45	−0.40	−0.04
Relative bias ($\times 10^{-3}$)	−0.22	4.20	−3.79	−1.23	−4.40	−4.01	−0.32
Standard error ($\times 10^{-3}$)	6.29	3.34	0.23	0.39	2.23	2.29	0.60
95% Coverage probability	94.6%	92.0%	94.2%	96.2%	96.2%	97.0%	94.4%

aFor ω_s and $\tilde{\omega}_s$, only estimates of selected parameters are presented, due to the large number of parameters. Estimates of the other parameters not shown here yield similar results.

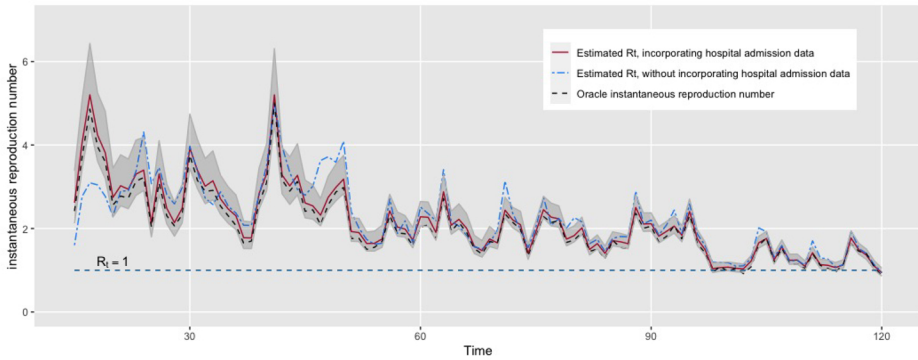

Fig. 3. Estimation of the instantaneous reproduction number when the daily new infections were reported with 0% to 30% under-reporting rates. The dotted line stands for the oracle instantaneous reproduction number. The solid line and the shaded area stand for the estimates and its corresponding Bootstrapping confidence interval (90% confidence level) using the proposed method. The dashed line stands for the estimates using a comparison method (Shi et al. 2022).

daily hospital admission data were accurate, ensuring that all hospitalized patients were reported and documented in a timely manner. Using the oracle reproduction number, we simulated true infections, underreporting percentages, and hospitalizations for each replication. The proposed method produced results similar to the correctly specified model. For example, the mean point estimates for β in the mis-specified model $(-0.0206, -0.1239)$ were nearly identical to those in the correctly specified model $(-0.02, -0.125)$, though the standard errors were higher in the mis-specified scenario $(0.003, 0.009)$ compared to the correctly specified model $(0.0003, 0.0004)$. Compared to a recent measurement error model (Shi et al. 2022), the proposed method performed slightly better, providing more accurate estimates of the reproduction number (Fig. 3).

5. APPLICATION TO COVID-19 DATA

We applied the proposed method to COVID-19 data from January 1 to June 30, 2021, for four major metropolitan counties: Miami-Dade, FL; New York, NY; Cook (Chicago), IL; and Wayne (Detroit), MI. These counties represent four major metropolitan areas in the Southeast coast, Tri-State area, and Great Lakes region of the United States. County-level data on daily new infections and hospitalizations due to COVID-19 were obtained from the National Healthcare Safety Network (NHSN) database. Additionally, two county-level risk factors were sourced: daily social distancing practices, indicated by the percentage change in visits to nonessential businesses (from Unacast), and wet-bulb temperature (from the National Oceanic and Atmospheric Administration). These risk factors were selected based on a thorough review of the literature that has

extensively examined local factors influencing COVID-19 transmission (Rubin et al. 2020; Talic et al. 2021; Weaver et al. 2022). Our objectives were to estimate the daily reproduction number R_t, infectiousness function ω_s, and hospitalization propensity $\tilde{\omega}_s$, and to assess the association between county-level factors and disease transmission. These findings aim to inform public health policies and resource allocation.

When fitting the proposed model, we allowed hospitalization propensity to vary by county to account for differences in healthcare resources and access. Due to limitations in US COVID-19 surveillance data, which recorded incidence at the time of disease diagnosis rather than actual infection, and considering that the exact infection times are rarely known, we interpreted the infectiousness function and hospitalization propensity estimated from this data based on diagnosis time rather than infection time. Specifically, $\tilde{\omega}-1, c$ and $\tilde{\omega}s, c$ represent the probabilities of never being hospitalized and being hospitalized on the s-th day after diagnosis, respectively, for each county $c \in 1, 2, 3, 4$. This approach captures county-level heterogeneity, while ω_s reflects infectiousness on the s-th day post-diagnosis. Bootstrap confidence intervals (CIs) were computed using the block approach (Bühlmann and Künsch 1999). The time series structure (2.4) was selected via AIC from a family of autoregressive models with exogenous terms, and the lengths of the infectiousness function and hospitalization propensity (η and $\tilde{\eta}$) were determined by maximizing the composite likelihood.

The final model selected was the same AR(1) structure used in the simulation study, with η and $\tilde{\eta}$ estimated at 22 and 4, respectively. The estimated effect sizes were 0.1539 (95% CI: 0.1537 to 0.1541) for social distancing and -7.3×10^{-4} (95% CI: -8.3×10^{-4} to -6.4×10^{-4})) for temperature, respectively. These results suggested that lack of social distancing was a strong risk factor for elevated disease transmission during the study period. For example, a 50% reduction in the frequency of visiting non-essential businesses was estimated to reduce R_t by an average of 7.4%. On the contrary, temperature exerted only a minor effect on disease transmission. The estimated county-level R_t exhibited a similar trend during the study period among all four counties, as depicted in Fig. S8. A decrease in disease transmission was observed since April 2021, which, in this dataset, was largely attributed to a reduction in social distancing value. Figure S9 illustrates the estimated infectiousness function. Its shape resembled the probability density function of either a Gamma or Weibull distribution (Fig. S9a), with nearly two-thirds of secondary infections being diagnosed within the first week after the diagnosis of the infectors (Fig. S9b). If we assume that the duration between infection and diagnosis is roughly the same for both infectors and secondary infections, our finding suggests that timely testing and a subsequent week-long quarantine of infected individuals can significantly mitigate disease transmission. The estimated hospitalization propensity for each county is presented in Fig. 4. Given the varied access to healthcare, hospitalization propensity diverged across locations. New York exhibited the highest propensity for hospitalization post-diagnosis (just under 20%), compared to about 10% in Miami. Hospitalization on the first day of diagnosis was also highest in New York (15.5%) but below 5% in Miami and Cook (Fig. 4a). Across counties, most hospitalizations occurred within 4 d of diagnosis, with an average of 1 d from diagnosis to admission. Assuming a 48-h delay from symptom onset to diagnosis, most hospitalizations occurred within 6 d of symptom onset, with a mean time to admission of approximately 3 d. These findings align with prior studies. For example, a CDC report estimated that 20.9% of U.S. COVID-19 patients were hospitalized before March 28, 2020 (Chow et al. 2020). Zhang et al. (2020) found a decrease in mean time from symptom onset to admission in Hubei, China, from 4.4 d (Dec 24–Jan 27, 2019) to 2.6 d (Jan 28–Feb 17, 2020). Traditionally, studying such durations requires epidemiological studies with contact tracing data, which poses a high requirement for the US disease surveillance system. Our results demonstrate that, even without extensive contact tracing, U.S. surveillance data can be effectively leveraged to estimate critical parameters for hospital planning and outbreak response.

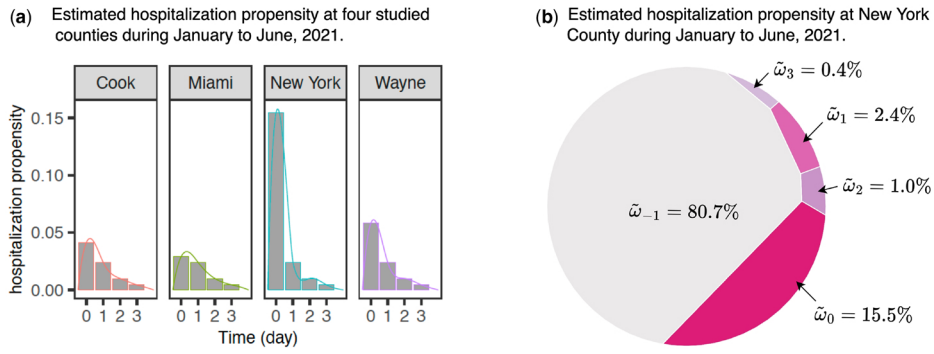

Fig. 4. Visualization of the estimated hospitalization propensity, $\tilde{\omega}_s$, for COVID-19 in the four studied counties during February to May 2021. In panel a), the bar plot shows the estimated value of $\tilde{\omega}_s$ $1 \leq s \leq 4$, at the four studied counties, while the solid lines represent the estimated propensity after fitting a cubic smoothing spline. In panel b), the estimated $\tilde{\omega}_s$ values at New York County are visualized using a Voronoi diagram to compare the magnitude of each $\tilde{\omega}_s$. Most of the hospitalized patients were admitted to a hospital on the day of infection.

6. DISCUSSION

This study introduces a new model and estimation procedure that extends traditional TSI models by incorporating data augmentation for both incidence and hospitalization data, enabling more accurate estimation of the instantaneous reproduction number, especially when hospitalization data is more reliable than incidence data. Our model also facilitates estimating hospitalization propensity, previously achievable only through contact tracing, and assessing the association between risk factors and transmission dynamics. The model is broadly applicable where both incidence and hospitalization data are available, particularly when incidence data quality is low.

The proposed method offers several extension opportunities. First, it can incorporate additional data, such as daily death counts, PCR tests, and serological data, to improve parameter accuracy and model death counts. Second, spatial correlations in disease transmission could be added using traffic data to enhance model efficiency. Third, the model can accommodate evolving pathogen dynamics and immunity changes, enabling analysis of new virus variants and immunity impacts. Fourth, it could be adjusted to handle overdispersion in infection counts by using distributions like the negative binomial, as illustrated in the Supplementary Materials. Finally, while the algorithm's runtime is manageable [1.7 to 6.5 min compared to 0.06 min in Cori et al. (2013)], further optimization could improve efficiency.

In addition to these extensions, the selection of appropriate risk factors is critical for applying our method effectively. In our study, we identified relevant factors through a systematic literature review and consultations with domain experts, focusing on those that significantly influence COVID-19 transmission dynamics. This approach ensured that the model captures local variations in transmission and improves the accuracy of R_t estimation. Including irrelevant variables could introduce noise and bias, undermining the model's performance. To assist users in selecting relevant factors in practice, we propose the following practical guidelines: 1) conduct a comprehensive literature review and consult with domain experts to identify strong candidate risk factors; 2)ensure the availability of reliable data and assess the correlation between potential risk factors and R_t; 3)void multicollinearity and focus on variables that are most relevant to R_t.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at *Biostatistics Journal* online.

FUNDING

Funding for the project was provided by the National Institutes of Health under award R01HD099348 and Centers for Disease Control and Prevention under award U01CK000674.

CONFLICT OF INTEREST

None declared.

DATA AVAILABILITY

Software in the form of R code, together with a sample example for the simulation study and complete documentation for the COVID-19 application is available on https://github.com/Jiasheng-Shi/Infectious-Disease-Hospitalization. Additional requests for implementation can be directed to the corresponding author at jing14@pennmedicine.upenn.edu.

REFERENCES

Abbott S, Thompson J, Funk S, Hellewell J. 2021. EpiNow2: estimate real-time case counts and time-varying epidemiological parameters. R package version 1.3.2. Comprehensive R Archive Network (CRAN). Vienna, Austria.

Ali ST, Wang L, Lau EH, Xu X-K, Du Z, Wu Y, Leung GM, Cowling BJ. 2020. Serial interval of SARS-CoV-2 was shortened over time by nonpharmaceutical interventions. Science. 369:1106–1109.

Athreya KB, Ney PE, Ney PE. 2004. Branching processes. Courier Corporation, North Chelmsford, MA, USA.

Aylward B, Barboza P, Bawo L, Bertherat E, Bilivogui P, Blake I, Brennan R, Briand S, Chakauya JM, Chitala K, et al; WHO Ebola Response Team. 2014. Ebola virus disease in west africa—the first 9 months of the epidemic and forward projections. N Engl J Med. 371:1481–1495.

Bühlmann P, Künsch HR. 1999. Block length selection in the bootstrap for time series. Comput Stat Data Anal. 31:295–310.

Chen D, Lau Y-C, Xu X-K, Wang L, Du Z, Tsang TK, Wu P, Lau EHY, Wallinga J, Cowling BJ, et al 2022. Inferring time-varying generation time, serial interval, and incubation period distributions for COVID-19. Nat Commun. 13:7727.

Chow N, Fleming-Dutra K, Gierke R, Hall A, Hughes M, Pilishvili T, Ritchey M. et al; Team, CDC Covid-19 Response, Team, CDC COVID-19 Response, Team, CDC COVID-19 Response. 2020. Preliminary estimates of the prevalence of selected underlying health conditions among patients with coronavirus disease 2019—United States, february 12–march 28, 2020. Morbidity Mortality Wkly Rep. 69:382–386.

Cori A. 2021. EpiEstim: estimate time varying reproduction numbers from epidemic curves. R Package Version 2.2-4. Comprehensive R Archive Network (CRAN). Vienna, Austria.

Cori A, Ferguson NM, Fraser C, Cauchemez S. 2013. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 178:1505–1512.

Davis RA, Dunsmuir WT, Streett SB. 2003. Observation-driven models for poisson counts. Biometrika. 90:777–790.

Davis RA, Dunsmuir WT, Wang Y. 1999. Modeling time series of count data. Stat Textb Monogr. 158:63–114.

Davis RA, Dunsmuir WT, Yin W. 2000. On autocorrelation in a poisson regression model. Biometrika. 87:491–505.

Deng Y, You C, Liu Y, Qin J, Zhou X-H. 2021. Estimation of incubation period and generation time based on observed length-biased epidemic cohort with censoring for covid-19 outbreak in China. Biometrics. 77:929–941.

Fraser C. 2007. Estimating individual and household reproduction numbers in an emerging epidemic. PLoS One. 2:e758.

Ge Y, Wu X, Zhang W, Wang X, Zhang D, Wang J, Liu H, Ren Z, Ruktanonchai NW, Ruktanonchai CW, et al 2023. Effects of public-health measures for zeroing out different sars-cov-2 variants. Nat Commun. 14:5270.

Gostic KM, McGough L, Baskerville EB, Abbott S, Joshi K, Tedijanto C, Kahn R, Niehus R, Hay JA, De Salazar PM, et al 2020. Practical considerations for measuring the effective reproductive number, R_t. PLoS Comput Biol. 16:e1008409.

Gressani O, Wallinga J, Althaus CL, Hens N, Faes C. 2022. Epilps: a fast and flexible bayesian tool for estimation of the time-varying reproduction number. PLoS Comput Biol. 18:e1010618.

Jewell NP. 2021. Statistical models for covid-19 incidence, cumulative prevalence, and R T. J Am Stat Assoc. 116:1578–1582.

Kaufmann H. 1987. Regression models for nonstationary categorical time series: asymptotic estimation theory. Ann Statist. 15:79–98.
Kermack W, Ogilvy McKendrick AG. 1927. A contribution to the mathematical theory of epidemics. Proc R Soc Lond Ser A Contain Papers Math Phys Character. 115:700–721.
Li Q, Guan X, Wu P, Wang X, Zhou L, Tong Y, Ren R, Leung KSM, Lau EHY, Wong JY, et al 2020. Early transmission dynamics in wuhan, China, of novel coronavirus–infected pneumonia. N Engl J Med. 382:1199–1207.
Lindsay B. 1988. Composite likelihood methods. Contemporary Math. 80:220–239.
Lloyd-Smith JO, Schreiber SJ, Kopp PE, Getz WM. 2005. Superspreading and the effect of individual variation on disease emergence. Nature. 438:355–359.
Nash RK, Nouvellet P, Cori A. 2022. Real-time estimation of the epidemic reproduction number: scoping review of the applications and challenges. PLOS Digit Health. 1:e0000052.
Pan A, Liu L, Wang C, Guo H, Hao X, Wang Q, Huang J, He N, Yu H, Lin X, et al 2020. Association of public health interventions with the epidemiology of the covid-19 outbreak in wuhan, China. JAMA. 323:1915–1923.
Quick C, Dey R, Lin X. 2021. Regression models for understanding covid-19 epidemic dynamics with incomplete data. J Am Stat Assoc. 116:1561–1577.
Ross R. 1916. An application of the theory of probabilities to the study of a priori pathometry—Part I. Proc R Soc Lond Ser A Contain Papers Math Phys Character. 92:204–230.
Ross R, Hudson HP. 1917a. An application of the theory of probabilities to the study of a priori pathometry—Part II. Proc R Soc Lond Ser A Contain Papers Math Phys Character. 93:212–225.
Ross R, Hudson HP. 1917b. An application of the theory of probabilities to the study of a priori pathometry—Part III. Proc R Soc Lond Ser A Contain Papers Math Phys Character. 93:225–240.
Rubin D, Huang J, Fisher BT, Gasparrini A, Tam V, Song L, Wang X, Kaufman J, Fitzpatrick K, Jain A, et al 2020. Association of social distancing, population density, and temperature with the instantaneous reproduction number of SARS-CoV-2 in counties across the United States. JAMA Netw Open. 3:e2016099.
Shi J, Morris JS, Rubin DM, Huang J. 2022. Robust modeling and inference of disease transmission using error-prone data with application to sars-cov-2. arXiv, arXiv:2212.08282, preprint: not peer reviewed.
Svensson AA. 2007. A note on generation times in epidemic models. Math Biosci. 208:300–311.
Talic S, Shah S, Wild H, Gasevic D, Maharaj A, Ademi Z, Li X, Xu W, Mesa-Eguiagaray I, Rostron J, et al 2021. Effectiveness of public health measures in reducing the incidence of COVID-19, SARS-CoV-2 transmission, and COVID-19 mortality: systematic review and meta-analysis. BMJ. 375:e068302.
Unacast 2020. Social distancing scoreboard. Unacast, Ashburn, VA, USA. Retrieved from https://www.unacast.com/post/the-unacast-social-distancing-scoreboard.
Varin C, Reid N, Firth D. 2011. An overview of composite likelihood methods. Stat Sin. 21:5–42.
Wallinga J, Teunis P. 2004. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. Am J Epidemiol. 160:509–516.
Weaver AK, Head JR, Gould CF, Carlton EJ, Remais JV. 2022. Environmental factors influencing COVID-19 incidence and severity. Annu Rev Public Health. 43:271–291.
Wu JT, Leung K, Bushman M, Kishore N, Niehus R, de Salazar PM, Cowling BJ, Lipsitch M, Leung GM. 2020. Estimating clinical severity of COVID-19 from the transmission dynamics in wuhan, China. Nat Med. 26:506–510.
Zeger SL, Qaqish B. 1988. Markov regression models for time series: a quasi-likelihood approach. Biometrics. 44:1019–1031.
Zhang J, Litvinova M, Wang W, Wang Y, Deng X, Chen X, Li M, Zheng W, Yi L, Chen X, et al 2020. Evolving epidemiology and transmission dynamics of coronavirus disease 2019 outside Hubei province, China: a descriptive and modelling study. Lancet Infect Dis. 20:793–802.