Journal of the American Statistical Association

December 2021 • Volume 116 • Number 536

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Regression Models for Understanding COVID-19 Epidemic Dynamics With Incomplete Data

Corbin Quick, Rounak Dey & Xihong Lin

To cite this article: Corbin Quick, Rounak Dey & Xihong Lin (2021) Regression Models for Understanding COVID-19 Epidemic Dynamics With Incomplete Data, Journal of the American Statistical Association, 116:536, 1561-1577, DOI: 10.1080/01621459.2021.2001339

To link to this article: https://doi.org/10.1080/01621459.2021.2001339

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 116, NO. 536, 1561–1577: Applications and Case Studies Discussion
https://doi.org/10.1080/01621459.2021.2001339

Taylor & Francis
Taylor & Francis Group

Regression Models for Understanding COVID-19 Epidemic Dynamics With Incomplete Data

Corbin Quick*[a], Rounak Dey*[a], and Xihong Lin[a,b]

[a]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA; [b]Department of Statistics, Faculty of Arts and Sciences, Harvard University, Cambridge, MA

ABSTRACT

Modeling infectious disease dynamics has been critical throughout the COVID-19 pandemic. Of particular interest are the incidence, prevalence, and effective reproductive number (R_t). Estimating these quantities is challenging due to under-ascertainment, unreliable reporting, and time lags between infection, onset, and testing. We propose a Multilevel Epidemic Regression Model to Account for Incomplete Data (MERMAID) to jointly estimate R_t, ascertainment rates, incidence, and prevalence over time in one or multiple regions. Specifically, MERMAID allows for a flexible regression model of R_t that can incorporate geographic and time-varying covariates. To account for under-ascertainment, we (a) model the ascertainment probability over time as a function of testing metrics and (b) jointly model data on confirmed infections and population-based serological surveys. To account for delays between infection, onset, and reporting, we model stochastic lag times as missing data, and develop an EM algorithm to estimate the model parameters. We evaluate the performance of MERMAID in simulation studies, and assess its robustness by conducting sensitivity analyses in a range of scenarios of model misspecifications. We apply the proposed method to analyze COVID-19 daily confirmed infection counts, PCR testing data, and serological survey data across the United States. Based on our model, we estimate an overall COVID-19 prevalence of 12.5% (ranging from 2.4% in Maine to 20.2% in New York) and an overall ascertainment rate of 45.5% (ranging from 22.5% in New York to 81.3% in Rhode Island) in the United States from March to December 2020. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received June 2021
Accepted October 2021

KEYWORDS
COVID-19 transmission; Effective reproductive number; EM algorithm; Epidemic model; Missing data; Prevalence; Serological studies; Under-ascertainment

1. Introduction

Coronavirus disease 2019 (COVID-19), caused by the Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2), has spread to over 227 countries across all habitable continents. Governments and public health agencies have implemented multifaceted measures to contain the spread of the virus, and are now working to distribute vaccines. Monitoring the real-time spread and predicting future trends have been important to inform policy, public guidance, and resource allocation. Mathematical and statistical epidemic modeling has played a central role in these efforts (Inglesby 2020; Hao et al. 2020; Gostic et al. 2020).

Several organizations have published real-time data on newly confirmed infections, tests conducted, and deaths. The available data on COVID-19 have presented several challenges for statistical inference (Jewell, Lewnard, and Jewell 2020; Roda et al. 2020; Bertozzi et al. 2020). We describe three of these challenges below, focusing on the 2020 data.

First, many SARS-CoV-2 infections are unascertained. In Wuhan, an estimated 87% of infections prior to March 2020 were unascertained (Hao et al. 2020). Underascertainment has persisted across the world and throughout the COVID-19 pandemic. For example, in the U.S. state of Massachusetts, the COVID Tracking Project reported that 130,900 infections were confirmed by September 2020 (COVIDTracking 2021), whereas a CDC seroprevalence study estimated 252,717 infections by that date (Bajema et al. 2020). Under-ascertainment is partly driven by individuals who experience few or no symptoms after being infected, and are therefore less likely to seek testing. Globally, an estimated 25% of SARS-CoV-2 infections are asymptomatic or mildly symptomatic (Alene et al. 2021). Also, insufficient testing capacity has caused under-ascertainment, particularly in the early stages of the pandemic. Moreover, ascertainment rates may vary across regions and time due to differences in testing availability and public awareness.

Second, there are time lags between infection, symptom onset, testing, and reporting. COVID-19 symptoms appear on average 5 days after exposure (He et al. 2020), and many infected individuals do not receive testing until days after initial symptom onset. Further delays between testing and reporting are sometimes evident, as noted in Schechtman (2021) and elsewhere. COVID-19 testing data also show cyclical weekday trends and oscillations, which may reflect biases in reporting, testing capacity, and other factors (Bergman et al. 2020).

Third, while a variety of polymerase chain reaction (PCR)-test-based data and serological-test-based data on COVID-19 incidence and prevalence are publicly available (e.g., COVIDTracking 2021; CDC 2021; COVIDTracking 2021; JHU-CSSE 2021; USAFacts 2021). The most widely analyzed

CONTACT Xihong Lin xlin@hsph.harvard.edu School of Public Health, Harvard University, Boston, MA 2115.

*Corbin Quick and Rounak Dey are co-first authors.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

COVID-19 data include confirmed and probable infections, hospitalizations and deaths, PCR tests, and antibody tests over time. Many of these variables show discrepancies across data sources; for example, Schechtman (2021) discussed differences between PCR test counts reported by federal and state sources. Other data sets provide complementary information, for example, positive and negative PCR test counts provide partial information about incidence and ascertainment, while periodic population-based antibody studies provide information about the prevalence up to a small number of time points (here defined as the prevalence of having any past SARS-CoV-19 infection throughout the study period). However, care is needed to account for uncertainty, sampling designs, and test characteristics across these data sources to make inferences about the spread of disease in the population.

The rate of new infections in a population depends crucially on the number of individuals that are currently infectious, the levels of immunity in the population, and the contagiousness of the disease. The latter two factors are encapsulated by the effective reproductive number R_t, defined as the expected number of secondary infections arising from a single infectious individual at time t. Each of these quantities is important to understand past trends and predict the future, and each is challenging to estimate for COVID-19. The fraction of the population that is infectious at a given time point is determined by the numbers of infections in previous weeks, as COVID-19 infectiousness is estimated to last 8–20 days following symptom onset (Wölfel et al. 2020; van Kampen et al. 2021; He et al. 2020). Absent vaccines, the fraction of the population that is immune is approximately the prevalence, as COVID-19 infection generally leads to immunity; however, recurrent infections have been reported (Iwasaki 2021), and could potentially become more frequent as new SARS-CoV-2 variants emerge (Murray and Piot 2021).

Estimating the incidence and prevalence of COVID-19 based on reported infections is challenging due to under-ascertainment (Manski and Molinari 2021). Several estimation procedures have been proposed using auxiliary datasets, for example, on influenza-like illness (ILI) and COVID-19 fatalities (Lu et al. 2020). However, data on ILI and COVID-19 fatalities also suffer from unreliable reporting, and are further complicated by variation in non-COVID-19 ILI disease outbreaks and fatality rates. Seroprevalence studies, which use antibody tests to detect past infections in random or convenience samples, provide another means to estimate the prevalence. Beginning July 2020, the CDC began conducting seroprevalence studies every 2 weeks for a few weeks in each U.S. state (Bajema et al. 2020; Havers et al. 2020). However, these studies also have several limitations. First, few time points are available (12–14 points before March 2021), and temporal resolution is limited due to specimens being collected across multiple weeks and aggregated. Second, they have limited statistical precision due to sample size, stratification, and imperfect test sensitivity and specificity. Third, they derive from convenience samples, which may lead to bias despite any adjustments for demographic composition and test characteristics.

The effective reproductive number R_t describes the expected number of new infections that arise per infected case at time t. As such, R_t depends on the transmissibility of the disease, the level of immunity in the population, and other possibly time-varying factors (Wallinga and Teunis 2004). The rate of new infections tends to increase over time if $R_t > 1$, and otherwise tends to decrease if $R_t < 1$. R_t is of particular interest for assessing the efficacy of non-pharmaceutical interventions in slowing the spread of the disease (Flaxman et al. 2020; Pei, Kandula, and Shaman 2020; Pan et al. 2020). Several epidemic methods have been proposed to estimate R_t (Wallinga and Teunis 2004; Bettencourt and Ribeiro 2008; Cori et al. 2013), and were recently compared in Gostic et al. (2020).

Previous methods to estimate R_t have several limitations. First, they do not explicitly account for under-ascertainment, and implicitly assume that all infections are observed. R_t estimates under this assumption are robust if ascertainment is constant over time; however, this is unlikely for COVID-19 due to increases in testing capacity over time and periodic shortages. Second, they model R_t at each time point discretely followed by moving averages over time, and do not allow a direct regression-based analysis of geographical and time-varying covariates, such as containment policies. Third, they do not directly account for delays between infection, onset, and testing, which can cause the estimated R_t curves to appear shifted forward in time, or smooth out the temporal variation of interest (Gostic et al. 2020). Statistical deconvolution can be applied to confirmed infection count time series as a preprocessing step before estimating R_t (discussed also in Gostic et al. 2020; Petermann and Wyler 2020; Miller et al. 2020). However, this strategy does not account for uncertainty due to stochastic time lags in the subsequent statistical inference.

Compartmental models provide an alternate framework for infectious disease dynamics. In this approach, individuals in a population are partitioned into discrete compartments at different time intervals, and the transition rates between compartments over time are specified by a set of differential equations. The classical compartmental model is the susceptible-exposed-infectious-recovered (SEIR) model (Anderson and May 1992). A number of Bayesian compartmental models have been proposed for COVID-19 by extending the SEIR model (Hao et al. 2020; Ndaïrou et al. 2020; Tian et al. 2021). These models incorporate additional compartments to account for unascertained infections and time lags between infection and testing. However, they often make restrictive assumptions on R_t and other transition parameters, most commonly using piece-wise constant functions over time. For example, Hao et al. (2020) used piece-wise constant functions with intervals of length 9–21 days. These compartmental models are difficult to extend to flexible regression models on covariates.

In this article, we propose a multi-level regression framework, MERMAID (Multilevel Epidemic Regression Model to Account for Incomplete Data), to model epidemic dynamics with incomplete data in one or multiple regions. MERMAID addresses key challenges modeling COVID-19 dynamics through four improvements. First, we model R_t using a flexible regression model, which can capture smooth trends over time, as well as explicit effects of geographic and time-varying covariates. Second, we model the probability of infection ascertainment as a function of time-varying covariates (e.g., numbers of PCR tests performed). Third, to calibrate baseline ascertainment and estimate prevalence, we incorporate serological survey data as an explicit term of the likelihood in addition to modeling confirmed infection counts. Fourth, we account for stochastic time lags between exposure, onset and reporting by treating the unobserved dates of infection as missing data, and assuming that discretized lag times follow a categorical distribution.

To estimate the model parameters, we develop an efficient EM algorithm, which avoids more computationally costly Markov Chain Monte Carlo procedures. We evaluate the performance of MERMAID using simulation studies, and study its robustness by conducting sensitivity analyses in a range of scenarios of misspecified models. We apply the proposed methods to analyze state-level COVID-19 daily confirmed infection counts, PCR testing data, and serological survey data across the U.S. in 2020. We estimate the time-varying reproductive numbers, the time-varying population prevalences, and the effects of state-level containment policies. For discussions of extending the methods to the analysis of the 2021 data, see the Discussion Section and the Rejoinder.

The remainder of the article is organized as follows. Section 2 describes the proposed model. Section 3 presents the procedures for maximum likelihood estimation of parameters and statistical inference under the proposed model using the EM algorithm. Section 4 provides simulation studies to assess the performance and robustness of our method under a variety of scenarios. Section 5 applies MERMAID to analyze state-level COVID-19 data in 2020, followed by discussions in Section 6.

2. Multilevel Epidemic Regression Model

We developed MERMAID (Multilevel Epidemic Regression Model to Account for Incomplete Data), a statistical framework to estimate epidemic dynamics (R_t, ascertainment, incidence, and prevalence) over time in one or multiple regions. MERMAID incorporates three sources of data (confirmed infections, serological surveys, and PCR testing metrics) and comprises four model components. First, we model the numbers of infections over time as conditionally Poisson variables. Second, we model the time lags between infection and potential confirmation by binning infection counts on each day by the length of time lag. Third, we model the numbers of confirmed infections as conditionally binomial variables, where success (ascertainment) probability is a function of testing metrics. Fourth, we model positive tests in serological surveys using binomial or hypergeometric distributions. In this section, we describe each model component given complete data. Estimation and inference procedures given observed (incomplete) data are described in Section 3.

2.1. Complete Data Likelihood

We use the following notations throughout the article. Let Y_{it} denote the number of newly infected individuals in region i on day t for regions $i = 1, 2, \ldots, \mathcal{R}$ and days $t = 1, \ldots, T_i$. We assume that the number of days between infection and confirmation (or reporting) for each individual is iid following a discrete distribution with support $0, 1, \ldots, m_A$. For infected individuals that are never confirmed (i.e., unascertained), this time lag can be interpreted as a counterfactual variable, which would have occurred if the individual had been confirmed. In other words, we treat ascertainment as a subsequent thinning process, independent of lag times.

We denote the number of individuals infected on day t and potentially confirmed on day $t + k$ in region i by A_{itk} for $k = 0, \ldots, m_A$, and define $\boldsymbol{A}_{it} = (A_{it0}, \ldots, A_{itm_A})$ which bins the individuals infected on day t by the day of confirmation, so $Y_{it} = \sum_{k=0}^{m_A} A_{itk}$. The total number of infected individuals that are potentially confirmed on day t is $M_{it} = \sum_{k=0}^{m_A} A_{i,t-k,k}$, out of which C_{it} infections are actually confirmed (ascertained) on day t. (We set $Y_{it'} := 0$ and $A_{i,t',k} := 0$ for all time-points $t' \leq 0$ before the start of the outbreak.) The remaining $U_{it} = M_{it} - C_{it}$ individuals are unascertained on day t. Finally, we use seroprevalence survey data to calibrate the ascertainment rate model. At seroprevalence survey periods $j = 1, 2, \ldots, J_i$, we assume that antibody tests were performed in a random sample of N_{ij} individuals in region i at time τ_{ij}, of which K_{ij} individuals tested positive.

The overall model can be described as follows:

New infections at time $t : Y_{it}$, where

$$\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT_i}) \sim F_{\boldsymbol{Y}}(\cdot; \boldsymbol{\theta}_Y, \boldsymbol{\phi})$$

Infections Y_{it} binned by time to potential confirmation at time $t, t+1, \ldots, t+m_A$

$$\boldsymbol{A}_{it} = (A_{it0}, \ldots, A_{itm_A}) | Y_{it} \sim F_{\boldsymbol{A}|\boldsymbol{Y}}(\cdot | Y_{it}; \boldsymbol{\phi})$$

Potentially confirmed infections at time t:

$$M_{it} = A_{it0} + A_{i,t-1,1} + \cdots + A_{i,t-m_A,m_A}$$

Confirmed infections at time $t : C_{it} | \boldsymbol{A}_i \sim F_{\boldsymbol{C}|\boldsymbol{A}}(\cdot | \boldsymbol{M}_i; \boldsymbol{\theta}_C)$;

Seroprevalence data: $K_{ij} | \boldsymbol{Y}_i \sim F_{\boldsymbol{K}|\boldsymbol{Y}}(\cdot | \boldsymbol{Y}_i; N_{ij})$,

where the unknown parameters are $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\boldsymbol{Y}}, \boldsymbol{\theta}_{\boldsymbol{C}})$, known (fixed) parameters are $\boldsymbol{\phi}$, and $F_{(\cdot|\cdot)}(\cdot)$ denotes the corresponding joint/conditional distribution. Figure 1 provides a schematic view of the MERMAID model where the infection potential Λ_{it} is a function of daily infections from past m_Λ days, and is a weighted average of $Y_{i,t-m_\Lambda}, \ldots, Y_{i,t-1}$ as defined in Section 2.2.1 (Cori et al. 2013). A toy numerical example illustrating these notations is given in Supplementary material S1 Table 1.

The complete data across the $\mathcal{R}$ regions are $(\boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{K})$, where $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{\mathcal{R}})$, $\boldsymbol{M}, \boldsymbol{C}, \boldsymbol{K}$ are defined similarly, and $\boldsymbol{A} = (\boldsymbol{A}_{11}, \ldots, \boldsymbol{A}_{1T_1}, \ldots, \boldsymbol{A}_{\mathcal{R}1}, \ldots, \boldsymbol{A}_{\mathcal{R}T_{\mathcal{R}}})$. As the total numbers of newly infected individuals $Y_{it} = \sum_{k=0}^{m_A} A_{itk}$ and potentially confirmed individuals $M_{it} = \sum_{k=0}^{m_A} A_{i,t-k,k}$ on each day are both deterministic linear functions of $\boldsymbol{A}$, we denote the complete data by $(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{K})$ for simplicity. The complete data likelihood for this model can be written as follows:

$$\begin{aligned} L(\boldsymbol{\theta} | \boldsymbol{A}, \boldsymbol{C}, \boldsymbol{K}) &= P(\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{K} | \boldsymbol{\theta}) \\ &= \underbrace{P(\boldsymbol{Y}(\boldsymbol{A}) | \boldsymbol{\theta}_Y)}_{\text{Transmission}} \times \underbrace{P(\boldsymbol{A} | \boldsymbol{Y}(\boldsymbol{A}); \boldsymbol{\phi})}_{\text{Time lag}} \\ &\quad \times \underbrace{P(\boldsymbol{C} | \boldsymbol{M}(\boldsymbol{A}); \boldsymbol{\theta}_C)}_{\text{Ascertainment}} \times \underbrace{P(\boldsymbol{K} | \boldsymbol{Y}(\boldsymbol{A}))}_{\text{Seroprevalence}} \end{aligned} \quad (1)$$

where we here write $\boldsymbol{Y}(\boldsymbol{A}) = \boldsymbol{Y}$ to emphasize that $\boldsymbol{Y}$ is a function of $\boldsymbol{A}$.

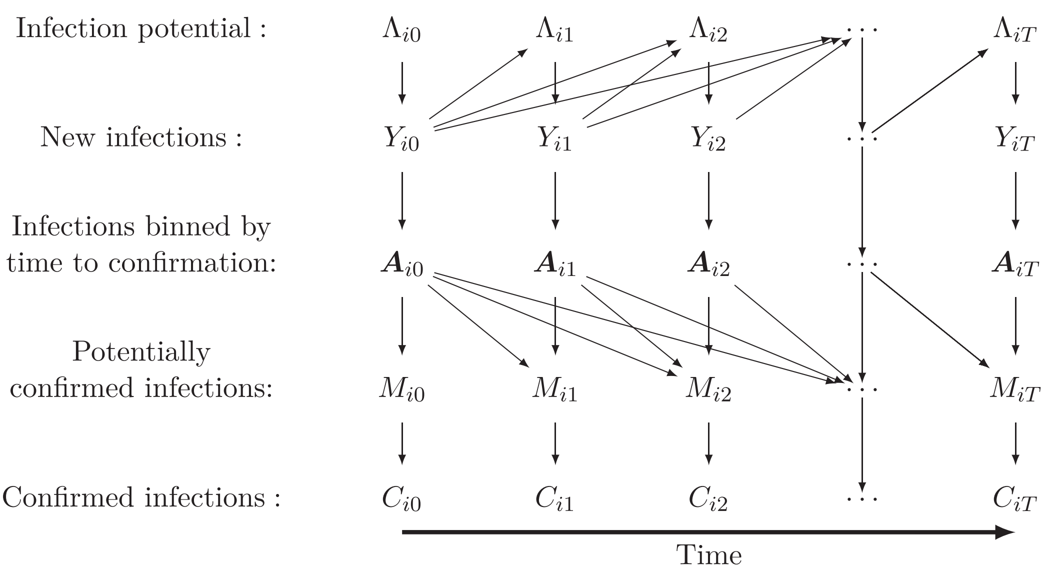

Figure 1. Simplified representation of the MERMAID framework. The number of infections Y_{it} in region i on day $t = 0, 1, ..., T$ depends on the current infection potential Λ_{it}, which is determined by the numbers of infections of the previous days. In practice, we assume that individuals are infectious for at most m_Λ days, so that Λ_{it} depends on $Y_{it'}$ only for $t - t' \leq m_\Lambda$ (not shown). New infections Y_{it} that arose on day t are binned by day that they are potentially confirmed; these binned infection counts are denoted by $\boldsymbol{A}_{it}$. The total number of individuals potentially confirmed on day t is denoted by M_{it}, and of these, only a subset of C_{it} individuals are confirmed, and the remaining $M_{it} - C_{it}$ individuals are labeled as unascertained. In practice, we assume that new infections are potentially confirmed at most m_A days after infection, so that M_{it} depends on $Y_{it'}$ only for $t - t' \leq m_A$ (not shown). For simplicity, serological survey outcomes and acquired immunity are not shown.

2.2. *Specification of Each Component of the Complete Data Likelihood*

In the following subsections, we describe each of the four components of the MERMAID model in detail for a single region. We assume all regions $i = 1, 2, ..., \mathcal{R}$ are independent, and so the complete data likelihood (1) can be written as a product of the single-region likelihoods.

2.2.1. *Modeling the Transmission Process $P(Y|\theta_Y)$*

Here, we describe a model of disease transmission adopting the notations of Cori et al. (2013). Let Y_{it} denote the number of new infections on day t in region i. Given the number of past infections $Y_{i0}, ..., Y_{i,t-1}$, we assume that

$$Y_{it}|Y_{i0}, ..., Y_{i,t-1} \sim \text{Poisson}(\Lambda_{it} R_{it}),$$

where the effective reproductive number R_{it} is defined as the expected total number of secondary infections that arise from a single primary infection given the level of exposure at time t in region i (Wallinga and Teunis 2004; Fraser 2007; Cori et al. 2013), and Λ_{it} is the infection potential in region i at time t. We define $\Lambda_{it} = \sum_{s=1}^{t} w_s Y_{i,t-s}$, where w_t is the probability that the serial interval (time between onset of a primary case and a secondary case that he or she infects) is equal to t days (between $t - 1/2$ and $t + 1/2$). This definition of Λ_{it} does not account for migration across regions, but could be extended to model between-region transmissions. As in Cori et al. (2013), we estimate the discretized serial interval weights $\{w_s\}_{s=0}^{\infty}$ using prior external data.

In practice, we assume that infectiousness lasts at most m_Λ days, hence $w_t = 0$ for $t > m_\Lambda$, and assume a fixed number of initial infections prior to the outbreak period ($t \leq 0$) within each region. Specifically, we define $\Lambda_{it} = \sum_{s=1}^{(t-1)} w_{t-s} Y_{is} + (1 - \sum_{s=1}^{t-1} w_s) y_{i\emptyset}$, where $y_{i\emptyset}$ can be interpreted as the average number of new infections or imported infections (assumed known) on or before time $t = 0$. Note that $(1 - \sum_{s=1}^{t-1} w_s) = 0$ for $t > m_\Lambda$, and therefore $y_{i\emptyset}$ only affects the first m_Λ time-points. This is similar to the Gamma(a, b) prior on R_t used by Cori et al. (2013), which effectively increases the infection potential by $1/b$ and the incidence by a.

The likelihood for daily infections over time in region i is given by

$$P(\boldsymbol{Y}_i) = \prod_{t=1}^{T_i} P(Y_{it}|Y_{i0}, ..., Y_{i,t-1}) = \prod_{t=1}^{T_i} \frac{(R_{it}\Lambda_{it})^{Y_{it}}}{Y_{it}!} e^{-R_{it}\Lambda_{it}},$$

where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT_i})^\top$. We model R_{it} using a (possibly semiparametric) log-linear regression model,

$$\log \boldsymbol{R}_i = \boldsymbol{a}_i^{(R)} + \mathbf{X}_i \boldsymbol{\beta}_i^{(R)}, \tag{2}$$

where $\boldsymbol{R}_i = (R_{i1}, \ldots, R_{iT_i})$, the offsets $\boldsymbol{a}_i^{(R)} = (a_{i1}^{(R)}, \ldots, a_{iT_i}^{(R)})$ are given by $a_{it}^{(R)} = \log(1 - p_{it})$, and p_{it} is the fraction of the population that is immune to infection at time t. We abuse notation by denoting $\log \boldsymbol{R}_i = (\log R_{i1}, \ldots, \log R_{iT_i})$. This mean model can include a B-spline basis matrix of time, and we can then write the linear predictor as a spline function of time t as, $\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i = \sum_l B_l(t)\beta_{il}$, where $B_l(t)$ denotes the lth B-spline basis. We use this approach in practice to model smooth trends over time. The independent variables $\mathbf{X}_i$ can further include time-varying or region-specific covariates that affect transmission rates, such as time-varying non-pharmaceutical interventions. This specification differs from Cori et al. (2013), where the effective reproductive number was estimated as a constant within a sliding window over time, and did not depend on covariates. A related maximum likelihood model for epidemic data was developed by Rojas et al. (2016) in which the probability of transmission was modeled using a piece-wise constant logistic regression equation, and subsequently used to estimate R_t.

2.2.2. Modeling Infection-Confirmation Lag Times $P(A|Y;\phi)$

Next, we model the time between infection and confirmation conditional on the total numbers of daily new infections (Gostic et al. 2020; Miller et al. 2020; Petermann and Wyler 2020). Here, confirmation refers to the time that a positive infection is reported, which we assume follows infection and disease testing. For mathematical convenience, we assign unascertained (unobserved) infections a counterfactual lag time, that is, the time at which they would have been ascertained had they been tested. We assume the number of individuals infected on day t with confirmation on day $t+k$ in region i is $A_{i,t,k}$, and,

$$A_{it0}, \ldots, A_{itm_A} | Y_{it} \sim \text{Multinomial}(Y_{it}, \boldsymbol{\phi}_i), \tag{3}$$

where $\boldsymbol{\phi}_i$ determines the distribution of days lag between infection and confirmation in region i. Therefore, the conditional likelihood of the lagged confirmations given the daily infections is,

$$P(\boldsymbol{A}_i|\boldsymbol{Y}_i) = \prod_{t=1}^{T_i} \binom{Y_{it}}{A_{it0}, \ldots, A_{itp}} \prod_{k=0}^{m_A} \phi_{itk}^{A_{itk}},$$

where $\boldsymbol{A}_i = (A_{i10}, \ldots, A_{iT_i0}, \ldots, A_{i1m_A}, \ldots, A_{iT_im_A})$ is the $T_i(m_A+1) \times 1$ vector of lagged confirmations. Miller et al. (2020) similarly proposed a method to account for delays in reporting using a categorical distribution; however, their method does not explicitly model the infection process or account for underascertainment.

2.2.3. Modeling Ascertainment Probabilities $P(C|M;\theta_C)$

The third component of the MERMAID likelihood accounts for underascertainment. Conditional on the number of potentially confirmed infections $M_{it} = \sum_{k=0}^{m_A} A_{i,t-k,k}$ on each day t in region i, the number of ascertained individuals C_{it} is assumed to follow a binomial distribution,

$$C_{it} \sim \text{Binomial}(M_{it}, \pi_{it}). \tag{4}$$

The corresponding conditional likelihood is given by

$$P(\boldsymbol{C}_i|\boldsymbol{M}_i) = \prod_{t=1}^{T_i} \binom{M_{it}}{C_{it}} \pi_{it}^{C_{it}} (1-\pi_{it})^{M_{it}-C_{it}}.$$

We model the probability of ascertainment using a logistic regression model

$$\text{logit}(\boldsymbol{\pi}_i) = \mathbf{Z}_i \boldsymbol{\beta}^{(\pi)}, \tag{5}$$

where $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iT_i})$. The covariates $\mathbf{Z}_i$ may include the numbers of tests performed, or the proportion of population being tested in region i, or other informative metrics for the ascertainment rate. The intercept parameter in Equation (5) is not identifiable from ascertained infection counts alone without further data or constraints. Here, the intercept is made identifiable by incorporating seroprevalence data in the likelihood.

2.2.4. Modeling Seroprevalence Survey Data $P(K|Y)$

The final component of the likelihood incorporates independent information on the prevalence of disease in the population from serological (seroprevalence) survey data. Suppose seroprevalence surveys were conducted on dates τ_{ij} for collection periods $j = 1, 2, \ldots, J_i$ in each region i. At collection period j, N_{ij} total antibody tests were performed and of them K_{ij} were positive. We assume that individuals were randomly selected from the population in region i on each of the survey dates, and that individuals test positive if and only if they have experienced an infection at any point in the past. Then, under random sampling,

$$K_{ij}|\boldsymbol{Y}_i \sim \text{Binomial}(N_{ij}, p_{ij}), \quad p_{ij} = \frac{1}{n_i} S_{ij} = \frac{1}{n_i} \sum_{t=0}^{\tau_{ij}} Y_{it}, \tag{6}$$

where n_i is the total population size of the region i, and $S_{ij} = \sum_{t=0}^{\tau_{ij}} Y_{it}$ is the total number of infections up to time t in region i. Therefore, the likelihood for this component of the model is,

$$P(\boldsymbol{K}_i|\boldsymbol{Y}_i) = \prod_{j=1}^{J_i} \binom{N_{ij}}{K_{ij}} p_{ij}^{K_{ij}} (1-p_{ij})^{N_{ij}-K_{ij}}$$

Alternatively, if individuals are tested without replacement, then the exact distribution of K_{ij} is hypergeometric (given n_i, S_{ij}, and N_{ij}), which is approximately binomial when n_i is large.

Given the numbers of infections $\boldsymbol{Y}_i$, the likelihood $P(\boldsymbol{K}_i|\boldsymbol{Y}_i)$ involves no unknown parameters. Rather, this component of the likelihood calibrates the conditional expectations $\mathbb{E}(\boldsymbol{Y}|\boldsymbol{C},\boldsymbol{K})$ that arise in the E step of the EM algorithm to estimate the parameters π_{it} and R_{it}.

When seroprevalence sample sizes are small relative to observed infection counts, the prevalence is nearly unidentifiable. Therefore, we multiply the seroprevalence log-likelihood by a constant, c_S, to increase its influence on the overall MERMAID log-likelihood. This weight has little effect on standard errors for $\hat{R}_{it}$, but can cause anti-conservative standard errors for ascertainment probabilities $\hat{\pi}_{it}$. We apply an approximate adjustment by multiplying the naive standard errors for $\hat{\pi}_{it}$ by a factor of $\sqrt{c_S}$. In real data applications, we set $c_S = 25$; R_{it} estimates were generally insensitive to the choice, while prevalence shows greater concordance with seroprevalence higher c_S values.

2.3. Observed Data Likelihood

From the available data sources, only the daily reported infection counts ($\boldsymbol{C}$) and numbers of positive antibody tests ($\boldsymbol{K}$) are observed. The remaining variables $(\boldsymbol{Y}, \boldsymbol{M}, \boldsymbol{A})$ in the complete-data log-likelihood (1) are unobserved. Let $\mathcal{D}_{\text{obs}} = (\boldsymbol{C}, \boldsymbol{K})$ denote the observed data, $\mathcal{D}_{\text{mis}} = \boldsymbol{A}$ denote the missing (unobserved) data, and $\mathcal{D} = (\mathcal{D}_{\text{obs}}, \mathcal{D}_{\text{mis}})$ the complete data. As the total number of infected individuals $Y_{it} = \sum_{k=0}^{m_A} A_{itk}$, and the total number of potentially observed individuals $M_{it} = \sum_{k=0}^{m_A} A_{i,t-k,k}$ can be written as deterministic functions of $\boldsymbol{A}_i$, the vectors $\boldsymbol{Y}$ and $\boldsymbol{M}$ are not shown explicitly in the missing data $\mathcal{D}_{\text{mis}}$. Then, the observed data log-likelihood is

$$\ell(\boldsymbol{\theta}|\mathcal{D}_{\text{obs}}) = \log \int L(\boldsymbol{\theta}|\boldsymbol{A}, \boldsymbol{C}, \boldsymbol{K}) d\boldsymbol{A},$$

which is not analytically tractable, and hence we use an EM algorithm to fit the model.

3. Estimation and Statistical Inference

This section describes estimation procedures for the unknown parameters $\boldsymbol{\theta}$ in MERMAID, which includes $\boldsymbol{\theta}_Y = \boldsymbol{\beta}^{(R)}$ for the effective reproductive numbers and $\boldsymbol{\theta}_C = \boldsymbol{\beta}^{(\pi)}$ for the probability of ascertainment. We assume that the remaining parameters, $\boldsymbol{\phi}$ for infection-reporting delay times in Equation (3) and $\{w_t\}_{t=1}^{m_\Lambda}$ for the serial interval distribution, are known and fixed. In practice, these fixed parameters can be specified based on previous literature or external data (discussed in Cori et al. 2013 for the serial interval and Petermann and Wyler 2020 for infection-reporting delays).

We estimate the model parameters using the EM algorithm, where we write the observed data log-likelihood as follows:

$$\begin{aligned}\ell(\boldsymbol{\theta}|\mathcal{D}_{\text{obs}}) = &\int P(\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}};\boldsymbol{\theta}')\log P(\mathcal{D}|\boldsymbol{\theta})d\mathcal{D}_{\text{mis}}\\ &-\int P(\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}};\boldsymbol{\theta}')\log P(\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}};\boldsymbol{\theta})d\mathcal{D}_{\text{mis}}\\ = &\,\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}') + H(\boldsymbol{\theta}|\boldsymbol{\theta}'),\end{aligned}$$

where $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}') := \int P(\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}};\boldsymbol{\theta}')\log P(\mathcal{D}|\boldsymbol{\theta})d\mathcal{D}_{\text{mis}}$ is the expected complete data log-likelihood where the expectation is taken with respect to the distribution of the missing data ($\mathcal{D}_{\text{mis}}$) given the observed data ($\mathcal{D}_{\text{obs}}$) evaluated at the parameter value $\boldsymbol{\theta}'$, and the second term $H(\boldsymbol{\theta}|\boldsymbol{\theta}') := -\int P(\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}};\boldsymbol{\theta}')\log P(\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}};\boldsymbol{\theta})d\mathcal{D}_{\text{mis}}$. In the traditional EM algorithm, the E step consists of calculating $\mathcal{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(c)})$ given the current estimate $\hat{\boldsymbol{\theta}}^{(c)}$, and the M step consists of maximizing $\mathcal{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(c)})$ with respect to $\boldsymbol{\theta}$ to obtain the next estimate $\hat{\boldsymbol{\theta}}^{(c+1)}$. Here, we use an accelerated variant of the EM algorithm in which the M step consists of a single Newton-type update using the expected complete data score vector $\boldsymbol{\mathcal{S}}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(c)}) = \partial\mathcal{Q}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(c)})/\partial\boldsymbol{\theta}$ and the observed data information matrix. Note that the expected complete data score vector $\boldsymbol{\mathcal{S}}(\boldsymbol{\theta}|\boldsymbol{\theta}')$ is equal to the observed-data score vector when evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}'$, as $H(\boldsymbol{\theta}|\boldsymbol{\theta}')$ is minimized (Lange 1995b).

Explicitly, the expected complete data score here is given by $\boldsymbol{\mathcal{S}} = (\boldsymbol{\mathcal{S}}_R, \boldsymbol{\mathcal{S}}_\pi)$ (we replace the notation $\boldsymbol{\mathcal{S}}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(c)})$ with $\boldsymbol{\mathcal{S}}$ where it is obvious to do so), where the components are,

$$\begin{aligned}\boldsymbol{\mathcal{S}}_R &= \textstyle\sum_{i=1}^{\mathcal{R}} \mathbf{X}_i^\top \left\{\mathbb{E}(\boldsymbol{Y}_i|\mathcal{D}_{\text{obs}}) - \boldsymbol{\mu}_i^{(R)}\right\},\\ \boldsymbol{\mathcal{S}}_\pi &= \textstyle\sum_{i=1}^{\mathcal{R}} \mathbf{Z}_i^\top \left(\boldsymbol{C}_i - \boldsymbol{\mu}_i^{(\pi)}\right).\end{aligned}$$

Here, the mean vectors are given by $\boldsymbol{\mu}_i^{(R)} = \left(\mu_{i1}^{(R)},\dots,\mu_{iT_i}^{(R)}\right)$ and $\boldsymbol{\mu}_i^{(\pi)} = \left(\mu_{i1}^{(\pi)},\dots,\mu_{iT_i}^{(\pi)}\right)$, where $\mu_{it}^{(R)} = R_{it}\mathbb{E}(\Lambda_{it}|\mathcal{D}_{\text{obs}})$ and $\mu_{it}^{(\pi)} = \pi_{it}\mathbb{E}(M_{it}|\mathcal{D}_{\text{obs}})$.

To find the maximum likelihood estimates, we iteratively calculate

$$\begin{aligned}\boldsymbol{\mathcal{S}}(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(c)}) &= \mathbb{E}_{\hat{\boldsymbol{\theta}}^{(c)}}\left\{\tfrac{\partial}{\partial\boldsymbol{\theta}}\ell\left(\boldsymbol{\theta}|\mathcal{D}\right)\middle|\mathcal{D}_{\text{obs}}\right\},\\ \hat{\boldsymbol{\theta}}^{(c+1)} &= \hat{\boldsymbol{\theta}}^{(c)} + \boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{obs}}}^{-1}\left(\hat{\boldsymbol{\theta}}^{(c)}\right)\boldsymbol{\mathcal{S}}\left(\hat{\boldsymbol{\theta}}^{(c)}\middle|\hat{\boldsymbol{\theta}}^{(c)}\right),\end{aligned}$$

and terminate when $\mathcal{Q}(\hat{\boldsymbol{\theta}}^{(c+1)}|\hat{\boldsymbol{\theta}}^{(c)}) - \mathcal{Q}(\hat{\boldsymbol{\theta}}^{(c)}|\hat{\boldsymbol{\theta}}^{(c)})$ falls bellow a specified threshold. To see that the above is a Newton-Raphson update, recall that the score equality implies $\boldsymbol{\mathcal{S}}(\hat{\boldsymbol{\theta}}^{(c)}|\hat{\boldsymbol{\theta}}^{(c)}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta}|\mathcal{D}_{\text{obs}})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}^{(c)}}$. Here, $\boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{obs}}}$ denotes the observed information matrix, which is given by Louis' formula (Louis 1982) as follows:

$$\boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{obs}}}(\boldsymbol{\theta}) = \boldsymbol{\mathcal{J}}_{\mathcal{D}}\left(\boldsymbol{\theta}\middle|\hat{\boldsymbol{\theta}}^{(c)}\right) - \boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}}}\left(\boldsymbol{\theta}\middle|\hat{\boldsymbol{\theta}}^{(c)}\right),$$

where $\boldsymbol{\mathcal{J}}_{\mathcal{D}}(\boldsymbol{\theta}|\boldsymbol{\theta}') = -\partial^2\mathcal{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}'\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top$ is the conditional expectation of the complete-data information matrix, and $\boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}}}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \text{var}_{\boldsymbol{\theta}'}\left\{\partial\ell(\boldsymbol{\theta}|\mathcal{D})/\partial\boldsymbol{\theta}\,|\mathcal{D}_{\text{obs}}\right\}$ is the attenuation of the information matrix due to the missing data. After convergence, $\boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{obs}}}$ is used to obtain the asymptotic standard errors for the model parameters. Explicit forms for these matrices are given in Supplementary material S1 Section C.

The proposed Newton–Raphson-type update is similar in spirit to the quasi-Newton acceleration of Lange (1995a) and the Newton-type EM algorithm of Oakes (1999), and differs primarily in the handling of $\boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}}}$. Lange (1995a) proposed a quasi-Newton acceleration in which the Hessian is approximated using gradients from previous iterations, and Oakes (1999) showed that $\boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{obs}}}(\boldsymbol{\theta})$ can be calculated using the derivatives of $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}')$. Here, we use an approximation to calculate the score and information matrix, which is described below.

Omitting terms that are constant with respect to $\boldsymbol{\theta}$, the complete data log-likelihood $\ell\left(\boldsymbol{\theta}|\mathcal{D}\right)$ is linear with respect to the missing data vector $\mathcal{D}_{\text{mis}} = \boldsymbol{A}$. Therefore, only the conditional expectation $\mathbb{E}_{\boldsymbol{\theta}'}(\boldsymbol{A}|\mathcal{D}_{\text{obs}})$ is required to calculate the score vector $\boldsymbol{\mathcal{S}}(\boldsymbol{\theta}|\boldsymbol{\theta}') = \partial\mathbb{E}_{\boldsymbol{\theta}'}\left\{\ell\left(\boldsymbol{\theta}|\mathcal{D}\right)|\mathcal{D}_{\text{obs}}\right\}/\partial\boldsymbol{\theta}$, and the conditional variance $\text{var}_{\boldsymbol{\theta}'}(\boldsymbol{A}|\mathcal{D}_{\text{obs}})$ is required to calculate $\boldsymbol{\mathcal{J}}_{\mathcal{D}_{\text{mis}}|\mathcal{D}_{\text{obs}}}(\boldsymbol{\theta}|\boldsymbol{\theta}')$. We approximate these moments using Laplace's method at the E-step.

Fully exponential Laplace approximations (Tierney, Kass, and Kadane 1989) have similarly been used to approximate intractable integrals in EM algorithms (Steele 1996; Rizopoulos, Verbeke, and Lesaffre 2009). Here, we instead use first-order approximations of the conditional moments, which are expected to perform adequately since each element of the missing data $A_{i,t,k}$ is marginally Poisson distributed with a relatively large rate. The Laplace approximation of $\boldsymbol{A}|\mathcal{D}_{\text{obs}}$ can be derived as a Taylor series approximation of the log of the conditional probability function,

$$\begin{aligned}\log P_{\boldsymbol{\theta}'}(\boldsymbol{A} = \boldsymbol{a}\,|\mathcal{D}_{\text{obs}}) \approx &\log P_{\boldsymbol{\theta}'}(\boldsymbol{A} = \boldsymbol{a}^*\,|\mathcal{D}_{\text{obs}})\\ &+ \frac{1}{2}(\boldsymbol{a}-\boldsymbol{a}^*)^\top\mathbf{H}_A(\boldsymbol{a}^*)(\boldsymbol{a}-\boldsymbol{a}^*),\end{aligned}$$

where $\boldsymbol{a}^* = \arg\max_{\boldsymbol{a}}\log P(\boldsymbol{A} = \boldsymbol{a}\,|\mathcal{D}_{\text{obs}})$, the gradient evaluated at $\boldsymbol{a}^*$ has vanished, and $\mathbf{H}_A = \partial^2\log P(\boldsymbol{A} = \boldsymbol{a}\,|\mathcal{D}_{\text{obs}})\big/\partial\boldsymbol{a}\partial\boldsymbol{a}^\top$ is the Hessian. This log-likelihood has the form of a multivariate normal distribution with mean $\boldsymbol{a}^*$ and covariance matrix $-\mathbf{H}_A(\boldsymbol{a}^*)$, which serve as first-order approximations to the conditional moments of $\boldsymbol{A}$ given $\mathcal{D}_{\text{obs}}$. Specifically, we approximate the conditional mean $\mathbb{E}_{\boldsymbol{\theta}'}(\boldsymbol{A}|\mathcal{D}_{\text{obs}})$ by $\boldsymbol{a}^*$ and the conditional covariance $\text{var}_{\boldsymbol{\theta}'}(\boldsymbol{A}|\mathcal{D}_{\text{obs}})$ by $-\mathbf{H}_A(\boldsymbol{a}^*)$ to calculate the score vector and information matrix as described above.

The conditional mode $\boldsymbol{a}^*$ does not have closed form; therefore, we find $\boldsymbol{a}^*$ at each E step iteration using a Newton-Raphson-type iterative algorithm. Convergence is generally attained quickly, as starting values can be recycled from the previous E step. Further details are given in Supplementary material S1 Section B.

4. Simulation Studies

We conducted simulation studies to evaluate the performance of MERMAID, including bias and calibration of model parameter estimates. This section is organized as follows: (i) data-generating procedures used across all simulation studies; (ii) simulation studies under correctly specified models, where the assumed model matches the data-generating model; and (iii) simulation studies with model misspecification, where one or more attribute of the data-generating model is misspecified in MERMAID. Additional simulations using a different but more realistic data generating model where R_t follows a smooth curve with respect to time, and ascertainment depends on the symptomatic/asymptomatic/uninfected status of the individuals, is presented in Supplementary material S1.

4.1. Simulation Procedures

Here, we describe procedures to simulate epidemic data. In this section, we describe the core simulation algorithm and define key parameters. In the subsequent sections, we state the specific parameter values and summarize the results from each simulation setting. The fixed model parameters are the regression coefficients in the reproductive and ascertainment regression models $\boldsymbol{\beta}_i^{(R)}$ and $\boldsymbol{\beta}_i^{(\pi)}$, respectively, for each region, and the serial interval infection and observation lag distribution weights are $\{w_t\}_{t=0}^{\infty}$ and $\{\phi_t\}_{t=0}^{\infty}$, respectively, which are truncated so that $w_t = 0$ for $t > m_\Lambda$ and $\phi_t = 0$ for $t > m_A$ for some thresholds m_Λ and m_A. The total population size n_i, total number of time points T, initial numbers of infections $y_\emptyset$, and reproductive and ascertainment regression covariate matrices $\mathbf{X}_i$ and $\mathbf{Z}_i$ respectively are also pre-specified and fixed constant across replicates within each simulation setting.

We simulated data for regions $i = 1, 2, ..., \mathcal{R}$ and time-points $t = 1, 2, ..., T$ as follows:

1. Set the infection potential $\Lambda_{it} = \sum_{s=1}^{t-1} w_s Y_{i,t-s} + I(t \leq m_\Lambda) \sum_{s=t}^{m_\Lambda} w_s y_\emptyset$.
2. Draw the number of new infections $Y_{it} \sim \text{Poisson}(R_{it}\Lambda_{it})$, where $R_{it} := \exp(\boldsymbol{x}_{it}^\top \boldsymbol{\beta}_i^{(R)})$.
3. Draw $\boldsymbol{A}_{it} \sim \text{Multinomial}(Y_{it}, \boldsymbol{\phi})$, where $A_{i,t,k}$ is the number of infections potentially confirmed on day $t + k$ ($k = 1, \ldots, m_A$) among the number of individuals infected on day t in region i, Y_{it}.
4. Calculate the number infections that are potentially confirmed on day t as $M_{it} = \sum_{s=0}^{m_A} A_{i,t-s,s}$.
5. Draw the number of infections that are actually confirmed as $C_{it} \sim \text{Binomial}(M_{it}, \pi_{it})$, where $\pi_{it} := \text{expit}(\boldsymbol{z}_{it}^\top \boldsymbol{\beta}_i^{(\pi)})$.

We simulated seroprevalence survey data for each region i and survey panels $j = 1, 2, ..., J_i$ by specifying the dates of survey panels τ_{ij} to be evenly spaced across all time points $t = 1, 2, ..., T_i$. We simulated positive serological tests K_{ij} in study panel j as $K_{ij} \sim \text{Hypergeometric}(n_i, N_i, \sum_{s=0}^{\tau_{ij}} Y_{is})$, where n_i is the population size, N_i is the seroprevalence sample size (which is constant across study panels), and $\sum_{s=0}^{\tau_{ij}} Y_{is}$ is the total number of infections that occurred by time τ_{ij} in region i.

4.2. Simulation Results Under Correctly Specified Models

To assess bias and calibration of R_{it} and π_{it} estimates from MERMAID under correctly specified models, we considered 4 simulation settings, varying the data-generating model for R_{it}. In each setting, we simulated epidemics lasting 1 year ($T_i = 365$ days) in a single region with a population size of $n_i = 8{,}000{,}000$ and initial infection potential $y_\emptyset = 50$ for 500 replicates. We assumed a serial interval distribution with mean 4.7 days and standard deviation of 2.9 days (Nishiura, Linton, and Akhmetzhanov 2020) truncated to 30 days, and an infection-reporting lag distribution as $\text{NegativeBinomial}(r = 5, \mu = 5)$ truncated to 21 days. We assumed seroprevalence studies were conducted at $J_i = 6$ evenly spaced time points with sample sizes $N_i = 80{,}000$ (1% of the population), and used a log-likelihood weight of $c_S = 25$ (matching the value used in real-data analysis).

We specified $\log R_{it}$ curves using either (a) a B-spline basis or (b) a piece-wise constant function. The specific values and curves are shown in Figure 2. We then specified the ascertainment probabilities as $\text{logit}(\pi_{it}) = \beta_{0\pi} + z_{it}\beta_\pi$, where $\beta_\pi = 0.05$ and $\beta_{0\pi}$ is chosen so that the mean value of $\text{logit}(\pi_{it})$ is 0, roughly 50% ascertainment rate on average. We simulated the testing rate covariate as $z_{it} = \sqrt{t} + \sum_{k=1}^{4} a_k \cos(b_k + c_k t) + \epsilon_{it}$ where each ϵ_{it} is iid normal with mean 0 and variance 5/2. This specification mimics the period trends and gradual increase in testing in the United States; the resulting curves are shown in Figure 3A.

First, we assessed R_{it} estimates from MERMAID (Figure 2). For comparison, we further fit Poisson generalized linear models (GLMs) using either (a) the raw reported daily infection counts (C_{it}) or (b) the reconvolved infection-counts (as described in Petermann and Wyler 2020), defined as $\tilde{C}_{it} = \sum_{s=0}^{m_A} \phi_s C_{i,t+s}$ where ϕ_s is the probability that an infection-reporting lag is s days, as proxies for the complete-date response Y_{it}. For the reconvolved infections, we used the last observation carried forward to fill missing values at the end of the epidemic. To assess the calibration of standard errors (SEs), we compared the mean estimated standard errors for $\hat{R}_{it}$ across simulation replicates, and the empirical SEs for $\hat{R}_{it}$ estimated using the standard deviation of $\hat{R}_{it}$ estimates across replicates. Estimates from MERMAID showed little bias, particularly when compared with the two proxy methods, while the proxy methods using reported infections and reconvolution infections gave biased estimates of R_t's. We calculated standard errors for R_{it} using the delta method, with $\hat{\text{SE}}(\hat{R}_{it})^2 = \hat{R}_{it}^2 \boldsymbol{x}_{it}^\top \hat{\text{var}}(\hat{\boldsymbol{\beta}}^{(R)})\boldsymbol{x}_{it}$. Standard errors for $\hat{R}_{it}$ were generally concordant with the standard deviations across replicates (Figure 2, panel B).

Second, we assessed estimates of the ascertainment probability π_{it}, which was fixed constant across the 4 settings as R_{it} was varied (Figure 3A). Ascertainment probability estimates from MERMAID showed low bias. We similarly calculated standard errors for $\hat{\pi}_{it}$ using the Delta method, with $\hat{\text{SE}}(\hat{\pi}_{it})^2 = c_S(\hat{\pi}_{it}(1-$

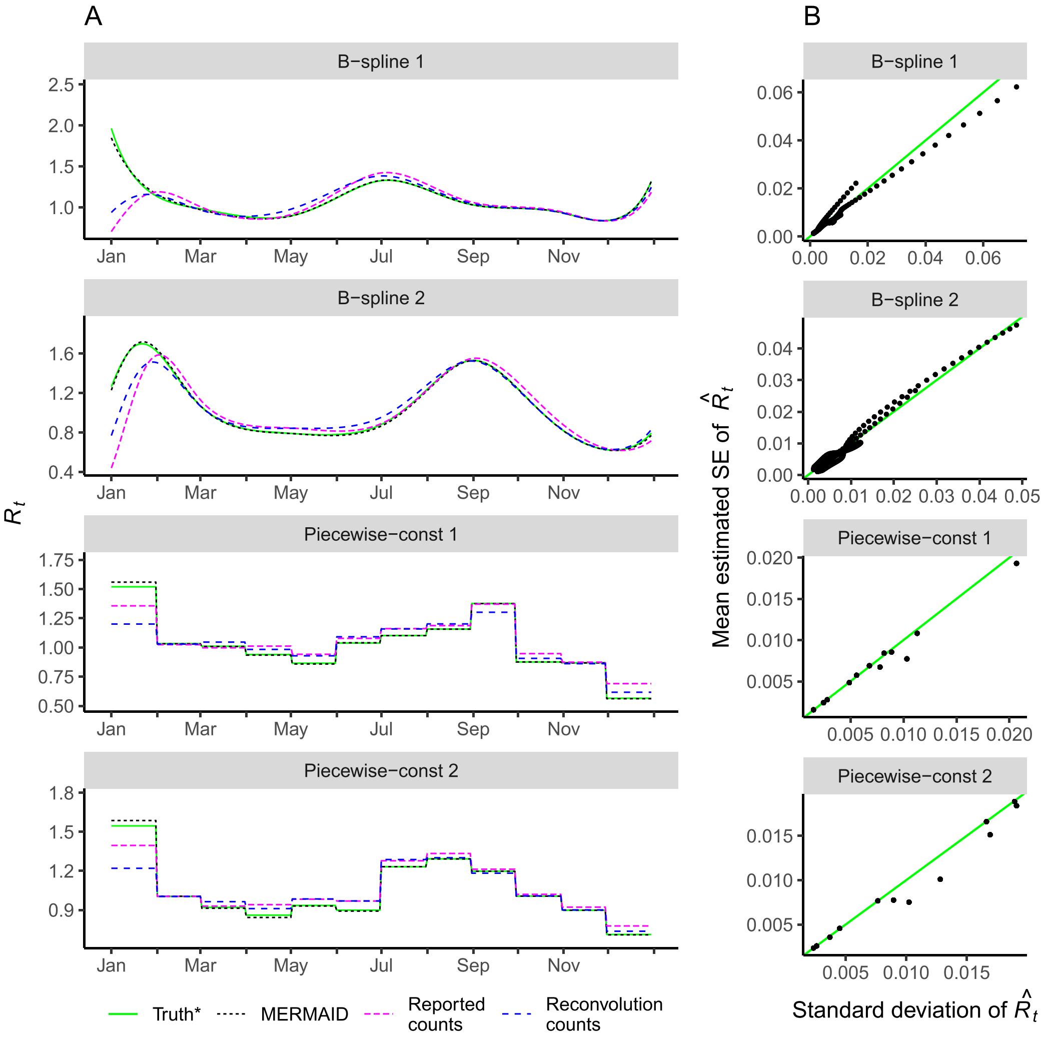

Figure 2. Estimates of R_{it} in 4 simulation scenarios each with 500 replicates under the correctly specified model, where the R_{it} is specified either using B-splines or piecewise constant functions. Panel A: Shown are curves for true R_{it} (solid green), and $\hat{R}_{it}$ estimated using MERMAID (black), a standard (misspecified) Poisson GLM with raw reported infection counts (magenta), or a Poisson GLM with reconvoluted infection counts (blue). Serosurveys were conducted on February 21, April 14, June 4, July 26, September 15, and November 6. Panel B: Comparison of mean estimated SE of $\hat{R}_{it}$ (y axis) and empirical SE of $\hat{R}_{it}$ for MERMAID, calculated as the SD of $\hat{R}_{it}$ (x axis) across replicates, in each scenario.

$\hat{\pi}_{it}))^2 \boldsymbol{z}_{it}^\top \hat{\text{var}}(\hat{\boldsymbol{\beta}}^{(\pi)})\boldsymbol{z}_{it}$, where the constant c_S adjusts for the seroprevalence likelihood weight. The adjusted SEs are conservative compared to the empirical SEs (Supplementary material S2).

Third, we assessed estimates of the predicted daily incidence (Figure 3B) and prevalence (Figure 3C). The daily incidence was defined as Y_{it}/n_i, where Y_{it} is the number of new infections on day t and n_i is the population size in region i, and the prevalence as $\sum_{s\leq t} Y_{it}/n_i$. The predicted values are calculated in MERMAID as $\hat{\mathbb{E}}(\cdot|\boldsymbol{K}_i, \boldsymbol{C}_i)$, and the standard errors of predictions using $\hat{\text{var}}(\boldsymbol{Y}_i|\boldsymbol{K}_i, \boldsymbol{C}_i)$. In this case, because $\boldsymbol{Y}_i$ varies across simulation replicates, we evaluated the calibration of standard errors by calculating the standard deviations of the centered predicted daily incidence, $\hat{\mathbb{E}}(Y_{it}/n_i|\boldsymbol{K}_i, \boldsymbol{C}_i) - Y_{it}/n_i$, across replicates and analogously for the prevalence. For the GLM models (using raw counts or reconvoluted counts, as described above), we plot the predicted values rather than the raw proxy counts. Both prevalence and daily incidence predictions from MERMAID showed little bias, and standard errors were well-calibrated (Figure 3 panels B and C; Supplementary material S2), whereas estimates based on raw or reconvoluted reported infection counts showed significant under-estimation due to failure to account for under-ascertainment.

4.3. *Simulation Results Under Misspecified Models*

We conducted additional simulations to evaluate the impact of model misspecification on the performance of MERMAID. First, we assessed the impact of the fixed parameters that characterize the distributions of the serial interval and infection-confirmation lag times, which may be uncertain in practice. Second, we assessed the impact of misspecification of the regression functions for the R_{it} and ascertainment probability π_{it} models.

We selected a single specification of the R_{it}, and simulated data under 8 settings by varying the data-generating distributions of infection-confirmation lags, which are assumed to follow Negative Binomial (r, μ), and the serial interval, which are assumed to follow a discretized Gamma distribution. For the lag distribution, we varied the stopping-time parameter $r \in 1, 999$,

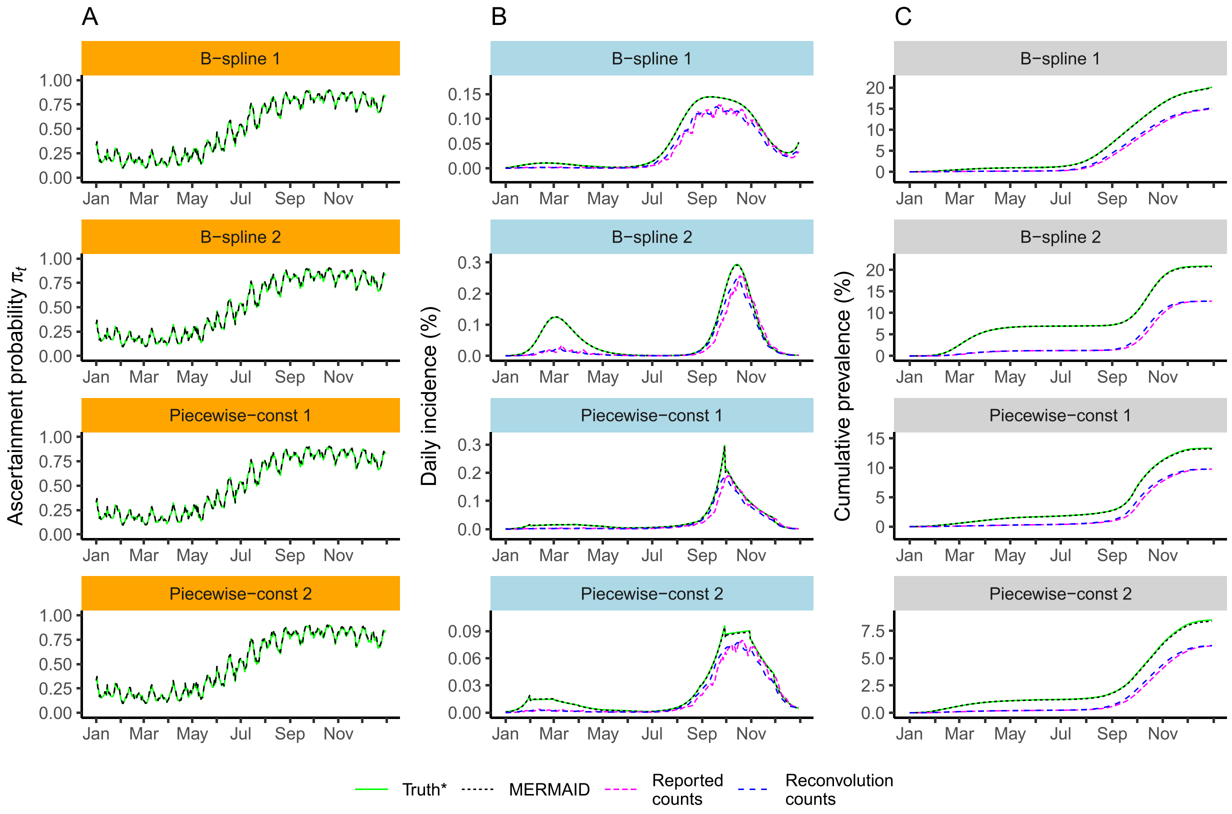

Figure 3. Ascertainment, daily incidence, and prevalence in 4 simulation settings each with 500 replicates. Ascertainment probabilities π_{it} are identical across the 4 simulation settings, while R_{it} is varied. In each panel, we compare true values (green) with estimates from MERMAID (black), which jointly estimates both π_{it} and R_{it}, and standard (misspecified) Poisson GLMs that ignore ascertainment and estimate R_{it} using either raw reported infection counts (magenta) or reconvoluted infection counts (blue). Serosurveys were conducted on February 21, April 14, June 4, July 26, September 15, and November 6. Panel A: True ascertainment probabilities (solid green) and estimated values from MERMAID (black) in each simulation setting. Panel B: Mean daily incidence (percentage of population infected on each day) across replicates. Panel C: Mean prevalence (percentage of population that has ever been infected) across replicates over time.

with the mean parameter $\mu = pr/(1-p) \equiv 5$ held constant. Here, a larger value of r corresponds to a larger variance of the lag times. For serial interval distribution, we varied the mean of the serial interval (3.7, 4.7, or 5.7), and the standard deviation of the serial interval (1.9, 2.9, or 3.9). For each setting, we fit MERMAID models to the simulated data and misspecified the lag distribution as NegativeBinomial$(r = 5, \mu = 5)$ (i.e., $p = 1/2$), and the serial interval as a discretized Gamma distribution with mean 4.7 and standard deviation 2.9 (as in the previous simulations). Results are shown in Supplementary material S2. In general, estimates were robust to misspecification of serial interval and lag; the mean of the serial interval had the greatest effect on the bias for R_{it} estimates. Ascertainment probabilities, prevalence, and incidence showed little bias under misspecification of the serial interval and lag distribution.

We also conducted simulations studies where the data are generated under a more realistic model where the probability of getting tested varies depending on whether the individuals are symptomatic, asymptomatic, or uninfected, and R_t follows an arbitrary smooth curve. The data generating model and simulation results are presented in Supplementary material S1 Section E. The results suggest that the estimates for R_t and ascertainment probabilities based on MERMAID are robust under moderate misspecification of the regression functions.

5. Data Application

We applied MERMAID to analyze COVID-19 transmission dynamics across US states in April-December 2020 using confirmed infection count, testing rate, and seroprevalence survey data from multiple sources. This section is organized as follows: First, we describe data sources and processing procedures. Second, we describe the R_t profile, incidence, and prevalence estimates in selected individual US states. Third, we describe changes in R_t associated with state containment policies across the US from the MERMAID analysis.

5.1. Data Sources and Processing Procedures

This subsection describes data sources and preprocessing procedures for (1) COVID-19 reported infections over time, (2) COVID-19 tests conducted over time, and (3) seroprevalence surveys.

5.1.1. COVID-19 Reported Infection Counts and PCR Tests

We obtained daily reported COVID-19 infections across US states from three sources: (1) the CDC data repository (CDC 2021), (2) the COVIDTracking project (COVIDTracking 2021), and (3) the USAFacts.org webpage (USAFacts 2021). We obtained daily COVID-19 testing data (numbers of positive and negative PCR and other test specimens or individuals tested) from 2 sources: (1) the COVID Electronic Laboratory Reporting Program (CELR), conducted by the US federal government and available from the HealthData.gov data repository (HealthData.gov 2021) and (2) the COVIDTracking project, which primarily uses data from state-level sources (COVIDTracking 2021; Schechtman 2021). We noted substantial discordance between data sources for both the numbers of confirmed infections and the numbers of tests performed (Supplementary material S3–4). For subsequent statistical modeling, we primarily used data from CELR, which appeared to provide proper reporting and fewer irregularities than other data sources. Because CELR reports only the numbers of positive PCR test specimens over time (and not the number of unique individuals), we constructed a consensus confirmed-infection time series by re-scaling the numbers of positive specimens reported by CELR. To that end, we estimated time-varying scaling factors (the number of unique confirmed infections per positive test specimen) in each state using confirmed infection counts from the other three data sources (CDC 2021; COVIDTracking 2021; USAFacts 2021) as a reference. We used this approach to ensure that confirmed infection counts are well-aligned with PCR test counts over time. A complete description of all data QC and preprocessing procedures is given in Supplementary material S1 Section F.

5.1.2. COVID-19 Seroprevalence Surveys

We obtained data from nationwide commercial laboratory seroprevalence surveys from the CDC website (Bajema et al. 2020). These studies performed antibody tests in pseudo-random convenience samples at multiple time points within each U.S. state. We calculated an adjusted positive antigen test count for each survey using the adjusted prevalence estimate reported by the CDC, in which observations were weighted to adjust for differences in the distribution of sex, ethnicity, and age between the seroprevalence sample and population based on the American Community Survey (Bajema et al. 2020). Sample sizes ranged from 83 to 2,553 (mean 985) within seroprevalence survey panels across states, and each state had 4-8 survey panels (mean 7.77) in the year 2020, with the earliest panel occurring on August 1. Specimen collection periods for each survey panel ranged from 1 to 22 days (median 14). Because the available data were aggregated across collection dates within each panel, we identified each survey panel with the midpoint of its start and end dates in our later analysis.

5.2. MERMAID Application to U.S. COVID-19 Data

5.2.1. Modeling U.S. State-Specific Epidemic Dynamics

Using the daily reported COVID-19 infections, total and positive COVID-19 tests conducted, and seroprevalence survey data as described in Section 5.1.2, we fit MERMAID models across all US states between March 15 and December 31, 2020. We modeled R_{it} in each state i by specifying the covariate matrix $\mathbf{X}_i$ as a cubic B-spline basis with 9 evenly spaced knots, and modeled the probability of ascertainment π_{it} by specifying $\mathbf{Z}_i$ to include the log-transformed numbers of PCR tests conducted as a time-varying covariate. We expected considerable differences in R_{it} trajectories between states due to differences in the time of initial outbreak and policy interventions, and therefore specified separate B-spline coefficients $\boldsymbol{\beta}_i^{(R)}$ for each state i. Similarly, we specified state-specific ascertainment coefficients $\boldsymbol{\beta}_i^{\pi}$.

Because all regression coefficients were region-specific, we fit each U.S. state model separately as an independent unit. We specified the serial distribution as a discretized Gamma (Cori et al. 2013) with a mean of 4.7 and standard deviation of 2.9, matching the estimates reported in Nishiura, Linton, and Akhmetzhanov (2020), truncated to $m_\Lambda = 30$ days. We specified the lag distribution as NegativeBinomial($r = 5, \mu = 5$) based on the median incubation period (time between infection and symptom onset) estimated at 5 days by Lauer et al. (2020), and truncated to $m_A = 21$ days. We note that the time from infection to confirmation may be longer than the incubation period due to testing and reporting delays, or shorter if individuals test positive after exposure but before symptom onset.

Results for the states of Michigan (Figure 4) and Texas (Figure 5) are shown here as examples; results for all US states are included in Supplementary material S5. Six states (Alaska, Montana, Wyoming, North Dakota, South Dakota, and Vermont) had insufficient data or failed to converge and were therefore excluded from subsequent analysis. Notably, large numbers of unascertained infections were estimated in March-April 2020 in many states, as expected. Across all states, the estimated effective reproductive number was highest at the beginning of the study period (mid-March 2020), and typically lowest in April-May 2020. In many states, MERMAID estimated a large early wave of unascertained SARS-CoV-2 infections during this period. Ascertainment estimates were generally higher in later waves, as PCR testing increased across the US.

5.2.2. Prevalence and Ascertainment Across the United States

Across the 44 U.S. states, we estimated that 12.45% of the overall population was infected by December 24, 2020 (model-based 95% CI: 12.42%–12.49%). The estimated prevalence varied substantially across states, ranging from 2.4% in Maine (model-based 95% CI: 2.3%–2.5%) to 20.2% in New York (model-based 95% CI: 20.0%–20.3%) (Figure 6, Panel B). Overall, we estimate that 45.4% of all infections between March and December 2020 were ascertained in the 44 U.S. states (model-based 95% CI: 45.3-45.5%). The estimated percentage ascertained also varied, ranging from 22.5% ascertained in New York (model-based 95% CI: 22.3%–22.7%) to 81.3% ascertained in Rhode Island (model-based 95% CI: 76.6%–86.7%) (Figure 6, Panel A). Visualizations of all state-specific prevalence and ascertainment estimates are given in the Supplementary material S5. We caution that these model-based CIs are sensitive to model assumptions, and do not capture all sources of uncertainty (e.g., in SI parameters, which were fixed).

Comparing across states, we observed a negative trend between the estimated prevalence, estimated as $\sum_t \hat{Y}_{it}/n_i$, and proportion of infections ascertained, estimated as $\sum_t C_{it}/\sum_t \hat{Y}_{it}$ (Figure 6, Panel C). In contrast, the longitudinal association between ascertainment rate and prevalence is positive, as both have increased over time due to simultaneous disease

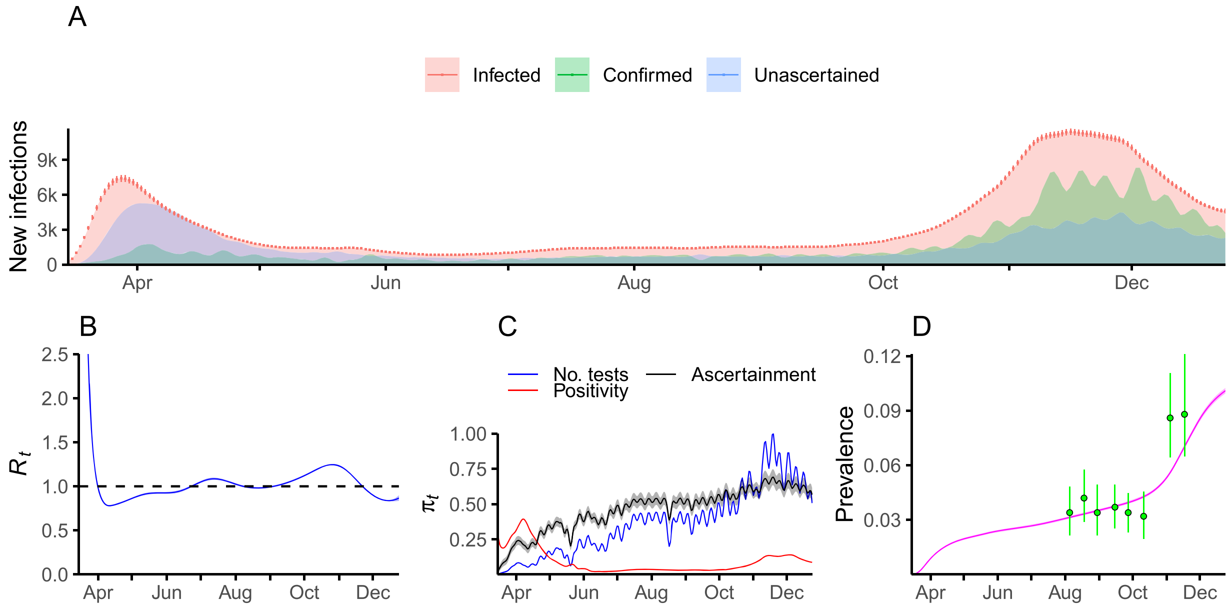

Figure 4. MERMAID analysis of COVID-19 epidemic dynamics in Michigan, March–December 2020. Panel A: Estimated total infections (red), estimated unascertained infections (blue), and confirmed infections (green) over time. Panel B: Estimated effective reproductive number over time. Panel C: Estimated ascertainment probability (black), percentage of positive PCR tests (red), and the total number of PCR tests scaled by its maximum value (blue) over time. Panel D: Estimated prevalence over time (magenta). Seroprevalence estimates and 95% confidence intervals reported from the CDC are shown in green.

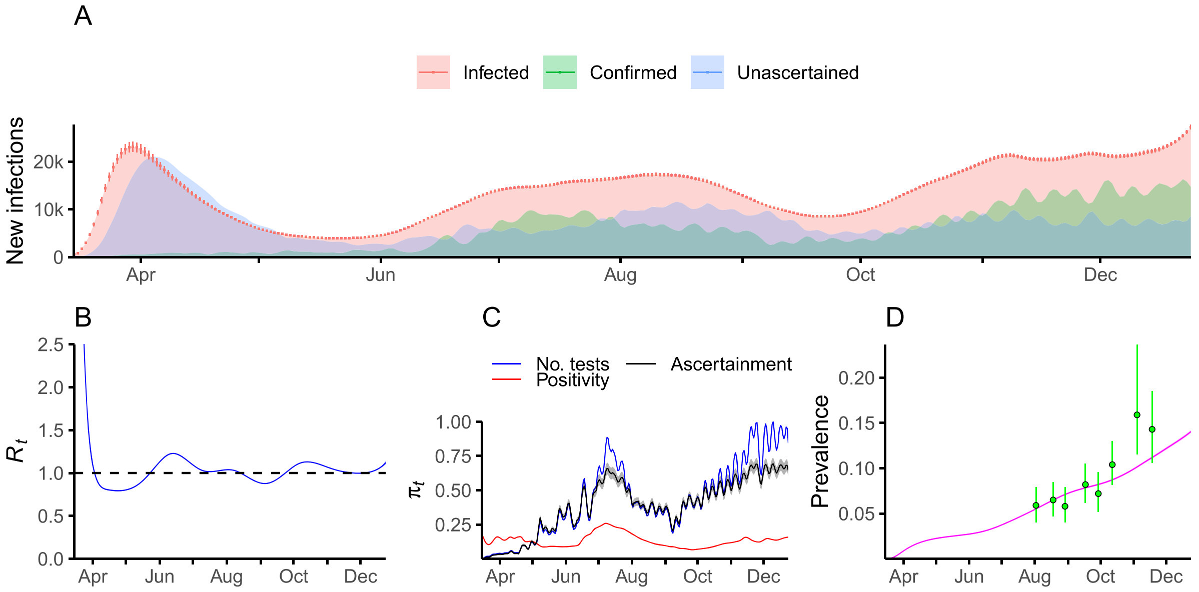

Figure 5. MERMAID analysis of COVID-19 epidemic dynamics in Texas, March–December 2020. Panel A: Estimated total infections (red), Estimated unascertained infections (blue), and confirmed infections (green) over time. Panel B: Estimated effective reproductive number over time. Panel C: Estimated ascertainment probability (black), percentage of positive PCR tests (red), and the total number of tests scaled by its maximum value (blue) over time. Panel D: Estimated prevalence over time (magenta). Seroprevalence estimates and 95% confidence intervals reported from the CDC are shown in green.

spread and increase in testing capacity (Figures 4 and 5, Panels C and D). However, across states, higher prevalence is associated with a lower mean ascertainment rate over time (defined as $\sum_{t=1}^{T_i} \hat{\pi}_{it}/T_i$; not shown) and a lower total proportion of infections ascertained (Figure 6, Panel C). Nevertheless, prevalence shows a strong positive trend with the proportion of the population that has been confirmed across states, as expected (Figure 6, Panel F).

We note three factors that may contribute to the negative trend between the overall ascertainment and prevalence across states (Figure 6, Panel C). First, the total percentage of infections ascertained is expected to be lower for states in which a larger fraction of infections occurred earlier in the pandemic, when testing capacity was more limited, for example, NY and NJ (Figure 6, panels D and E). Second, a negative trend between ascertainment and seroprevalence estimates could be caused by excess variability in seroprevalence estimates, particularly if the error variance is large relative to the actual variation in prevalence across states. In other words, a spurious negative correlation could arise from the fact that artificially increasing the seroprevalence estimate would drive the ascertainment estimate down and the prevalence estimate up, holding the numbers of

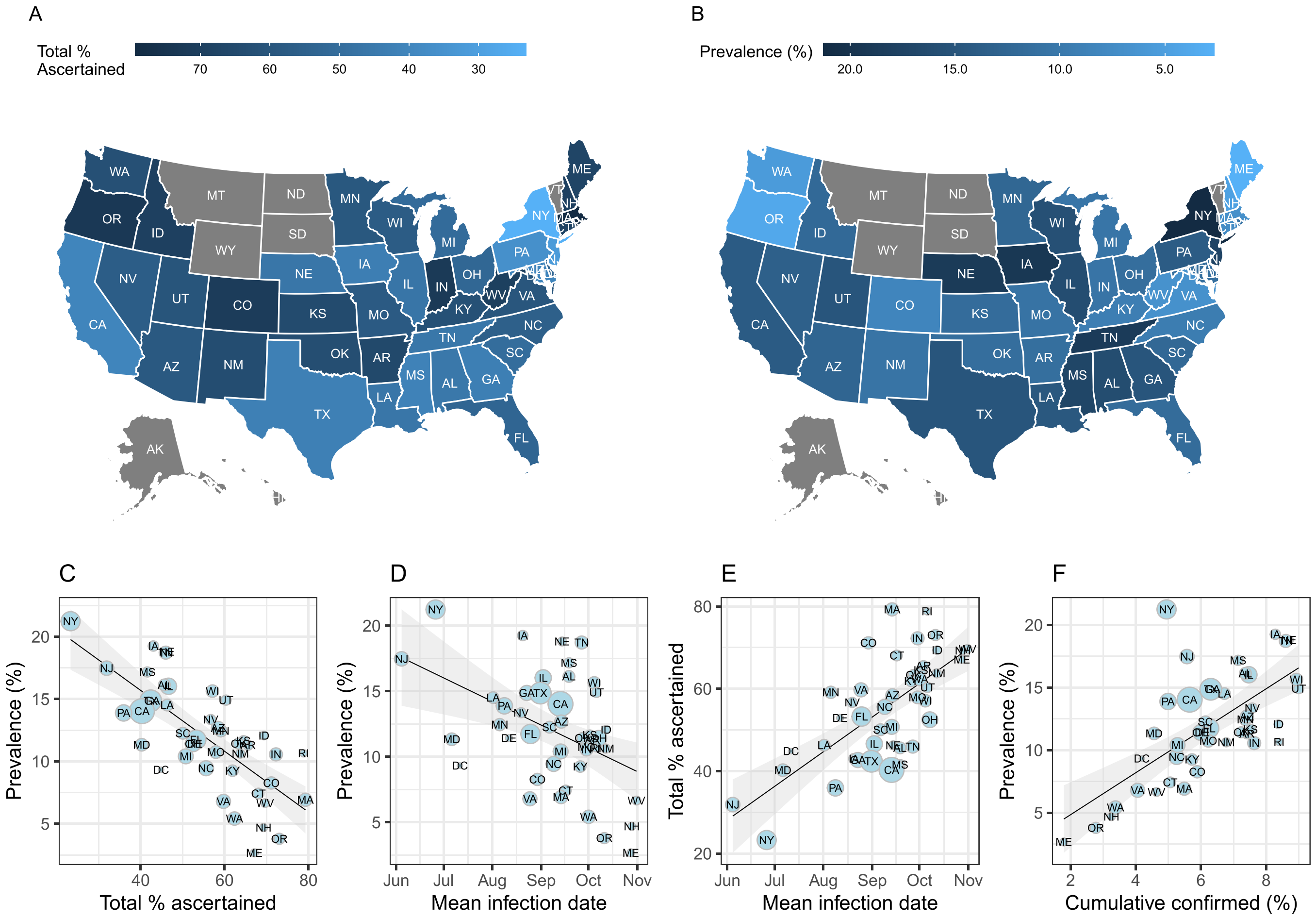

Figure 6. Estimated COVID-19 infection ascertainment fraction and prevalence across U.S. states as of December 2020. Panel A: Total percentage of infections estimated to be ascertained in each state, calculated as $\sum_{t=1}^{T_i} C_{it} / \sum_{t'=1}^{T_i} \hat{Y}_{it'}$. Panel B: Estimated prevalence in 2020, calculated as $\sum_{t=1}^{T_i} \hat{Y}_{it}/n_i$. States in gray in Panels A and B had insufficient data or failed to converge. Panel C: Estimated total percentage ascertained (x-axis) and prevalence (y-axis). Panel D: Estimated prevalence (y-axis) and mean infection date (x-axis), where mean infection date is calculated as $\sum_{t=1}^{T_i} t\hat{Y}_{it} / \sum_{t'=1}^{T_i} \hat{Y}_{it'}$. Panel E: Estimated total percentage ascertained (y-axis) and mean infection date (x-axis). Panel F: Estimated prevalence (y-axis) and the total percentage of the population ever confirmed positive (x-axis).

confirmed infections constant. Third, demand for PCR testing may be greater in states with greater prevalence, leading to more frequent and severe testing shortages.

5.3. Containment Policies and Effective Reproductive Numbers in the United States

We assessed differences in effective reproductive numbers R_t associated with containment policies over time across the US. We obtained data on containment policies from the Oxford Covid-19 Government Response Tracker (Hale et al. 2021). We considered 5 policy categories: (i) public transportation closures, (ii) facial covering mandates, (iii) gathering restrictions, (iv) stay-home orders, and (v) workplace closing. Within each category, greater policy levels indicate stricter enforcement. For example, stay-home policies include recommendations (Level 1), requirements with exceptions for exercise, grocery shopping, and essential trips (Level 2), and requirements with minimal exceptions (Level 3).

State containment policies are modeled as time-varying covariates, which vary between states and within states across time (Figure 7). In the previous section, we modeled $\log R_{it}$ using state-specific smooth functions of the form,

$$\log R_{it} = \beta_i(t) = \boldsymbol{\beta}_i^\top \boldsymbol{B}(t), \qquad (7)$$

where $\boldsymbol{B}(t)$ is a cubic B-spline basis with 9 knots. Time-varying policy covariate effects cannot be estimated simultaneously with state-specific time trends of this form. For example, under a fully saturated model for $\beta_i(t)$ (including the maximum number of estimable spline knots), the effects of time-varying covariates are clearly unidentifiable.

To quantify differences in $\log R_{it}$ associated with containment policies across all states regions, we considered the models with a shared baseline time trend of the form,

$$\log(R_{it}) = \beta_{0i} + \beta_B(t) + \sum_k \boldsymbol{X}_{ik}(t)^\top \boldsymbol{\beta}_{Xk} \qquad (8)$$

where β_{0i} is a state-specific intercept, $\beta_B(t)$ is a shared time trend unrelated to state policy, and $\boldsymbol{X}_{ik}(t)$ characterizes policies of category k in state i at time t. This model assumes that $\log(R_{it})$ curves are parallel across regions in the absence of containment policies, and that the temporal variation from this within states is due to changes in containment policies. We compared two specifications of the shared time trend $\beta_B(t)$: (1) $\beta_B(t) \equiv 0$,

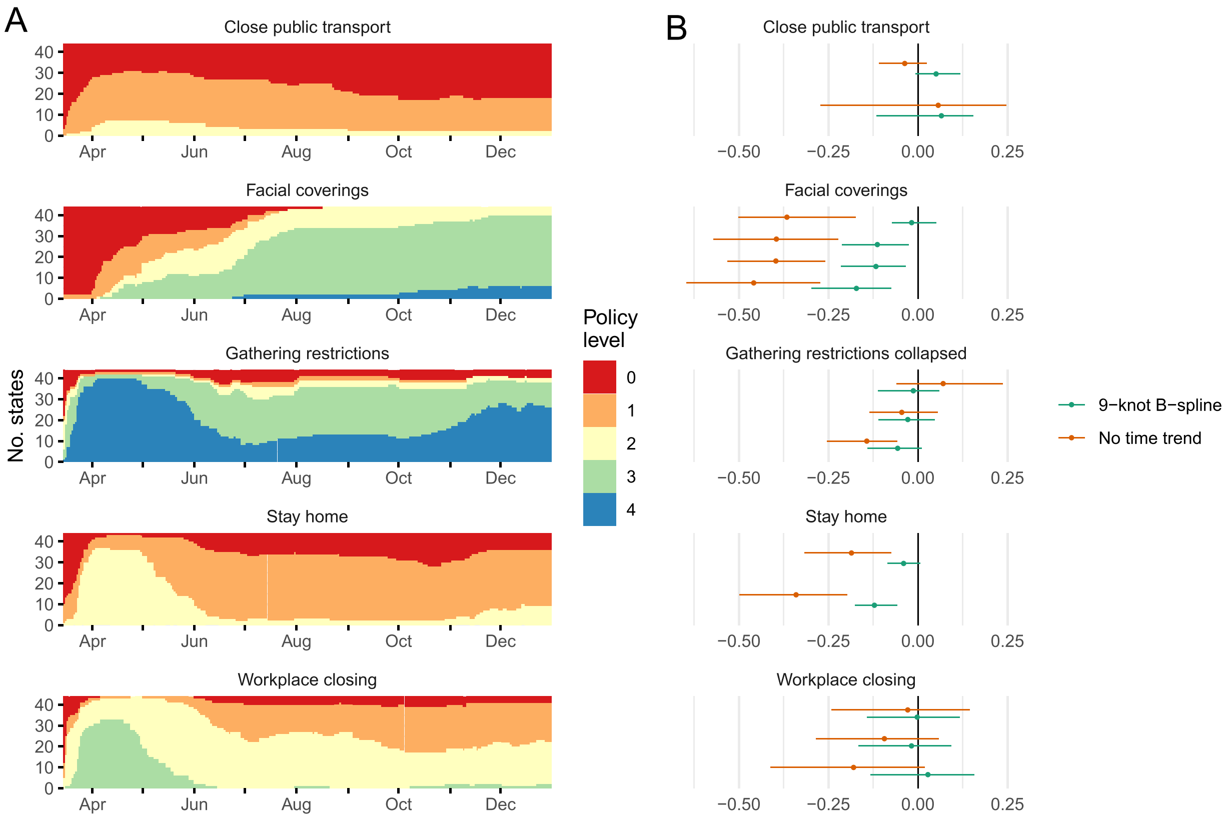

Figure 7. Differences in $\log R_{it}$ associated with non-pharmaceutical interventions. Panel A: Policy levels across U.S. states (*y*-axis) over time (*x*-axis) for six categories of containment policies as reported by the Oxford COVID-19 Government Response Tracker. In each panel, the y-axis denotes the number of states in which the policy is enforced at a specified level, where policy levels are indicated by color. Panel B: Differences in $\log R_{it}$ (x axis) associated with policy levels in each category estimated under model (8) with either (a) a shared U.S.-wide time trend specified as a cubic B-spline with 9 knots, or (b) no time shared trend, whereby all changes in $\log R_{it}$ across time are due to changes in policy. Shown are block bootstrap confidence intervals (Cameron, Gelbach, and Miller 2008), where each bootstrap block is a single state, and intervals correspond to the 2.5% and 97.5% percentiles of bootstrapped estimates across 1000 bootstrap samples.

whereby all changes in $\log R_{it}$ across time are attributable to changes in policy, and (2) $\beta_B(t)$ is modeled using a cubic B-spline basis with 9 knots, as in the state-specific model (7). While model (8) is restrictive, some modeling assumptions or constraints are necessary to estimate policy effects from the observed data.

We assumed that policy effects have the form $\boldsymbol{X}_{ik}(t)^\top \boldsymbol{\beta}_k = \sum_{j=1}^{m_k} X_{ikj}(t)\beta_{jk}$ where m_j is the number of levels for policy category k, and $X_{ikj}(t)$ is a dummy variable that specifies policy implementation. Here, $X_{ikj}(t) = 1$ for time points t when policy k level j is implemented in region i, and $X_{ikj}(t) = 0$ otherwise. We fit (8) using a computationally efficient two-stage approximate EM procedure described in Supplementary material S1 Section D.

Policy effect estimates for each of the 5 categories and 3 baseline time trend models are shown in Figure 7, where policy level 0 is the reference in each category. Greater policy levels were generally associated with greater decreases in $\log R_{it}$ within each policy category. This trend is expected, as greater policy levels signify broader application and stricter enforcement, and are expected to have greater adherence. Overall, stay-home and face-covering policies were associated with the largest decreases in $\log R_{it}$.

We note four factors that could mask or confound the effects of containment policies on $\log R_{it}$ in the U.S. First, the effects of policies that vary little across states over time are nearly unidentifiable in models that include state-specific intercepts and a shared time trend. Second, simultaneous deployment of multiple policy changes within states could mask the effects of policies that are partially redundant (e.g., gathering restrictions and stay-home orders) or more effective individually than in combination (e.g., face-covering mandates and stay-home orders). Third, strategic deployment of policies in anticipation of COVID-19 outbreaks (nonrandom assignment) could mask or apparently reverse the effect of policy (Hernán, Brumback, and Robins 2002). Given the recurring spread of COVID-19 throughout the U.S. in the time period we analyzed, this scenario appears less probable. Fourth, omitted temporal confounders (e.g., news reports and other communications that influence both preventive behaviors and policy) may introduce bias in policy effect estimates.

In addition, we note that the effects of policies may vary across time and states. For example, differences in baseline behavior across states (e.g., the proportion of the population that would wear a face-covering in the absence of any face-covering mandate) could contribute to differences in policy effects. Also, *de facto* enforcement of and adherence to policies may vary across regions and over time. Here, we assumed constant policy effects on $\log R_{it}$, as time- and state-specific effects are not identifiable from the data under our model. We therefore, interpret the policy effect estimates with caution.

6. Discussion

In this article, we presented a comprehensive and flexible statistical regression framework to estimate R_t, ascertainment rate, incidence, and prevalence of an infectious disease over time in one or more regions while accounting for under-ascertainment and reporting lags, by integrating data on reported infections, testing, and serological studies. The method, called MERMAID, solves four important challenges for modeling the dynamics of the COVID-19 epidemic from empirical data. First, it provides a principled framework to integrate confirmed infection data, PCR testing metrics, and serological data to estimate incidence and prevalence over time. Second, it accounts for temporal variation in under-ascertainment in reported infection counts by modeling the probability of ascertainment as a function of testing capacity metrics. Third, it accounts for the delay between exposure, symptom onset, and reporting by modeling stochastic time lags as missing data in an EM framework. Finally, it paves the way for regression-based analyses of epidemic dynamics by modeling the effective reproductive number R_t as an explicit function of covariates.

Through simulation studies, we showed that the EM estimation procedure performs well in estimating parameters of interest under a correctly specified model. We further showed that MERMAID is robust in realistic scenarios where the functional form of the ascertainment and R_t models are misspecified, or nuisance parameters characterizing the serial interval and lag distribution are incorrectly specified. While analytic standard errors (SEs) from MERMAID appeared well-calibrated in simulation studies with correctly specified or mildly misspecified models, the model-based SEs likely underestimate uncertainty in real data, where they often appear dubiously small. Thus, it is of interest to develop robust SEs for MERMAID and other maximum likelihood approaches for modeling epidemic data.

Our analysis of COVID-19 confirmed infections, PCR tests, and seroprevalence studies in the US highlighted difficulties establishing incidence and prevalence over time from public data sources, and a possible methodological resolution. Two common approaches to monitor COVID-19 incidence and prevalence are a) the confirmed infection counts, and b) the fraction of PCR tests that return positive. As expected, our analysis suggests that the percentage of the population that has been confirmed is substantially less than the prevalence in the US, while PCR test positivity is typically greater than the prevalence due to over-representation of infections. Seroprevalence studies provide an alternate approach to estimate prevalence, but have limited sample size and statistical precision, small numbers of time points, and possible biases due to convenience sampling. Here, we presented a statistical approach to estimate incidence and prevalence over time by integrating all three data sources.

Our analysis showed that containment policies are associated with substantially lower effective reproductive numbers across the US. As we discussed, the estimates of containment policy effects on R_t may be biased due to confounding or other misspecification. While policies have specific dates of implementation, human behavior (e.g., adherence to policy) can vary across time and regions. Human behavior and policy decisions may also be influenced by recent outbreaks and other events (e.g., via news reports), making it difficult to estimate causal effects. Data characterizing the efficacy of specific behaviors (e.g., face-covering and social distancing) for preventing transmission and adherence to existing recommendations are also therefore important to develop effective policies (e.g., Chu et al. 2020; Brooks, Butler, and Redfield 2020). Simulation studies under realistic epidemic models are also valuable tools to assess the potential impact of policies given adherence and efficacy (e.g., Adam 2020; Currie et al. 2020).

Our analysis relied on restrictive assumptions about PCR and antibody tests. First, we assumed that PCR tests have perfect sensitivity and specificity to detect infected individuals. In reality, PCR test sensitivity depends on the viral load, which varies continuously following exposure. This assumption may be reasonable, as the viral load is correlated with infectiousness (Lee et al. 2021). Second, we assumed that SARS-CoV-2 antibodies are detectable shortly after the infection, and remain detectable for up to 9 months after infection. In reality, SARS-CoV-2 antibody response varies across individuals, may be detectable only after 7-14 days following infection, and may decrease over time. These factors could be modeled using stochastic time lags as we used to account for the delays between infection and reporting, and weighting past infection counts as used to calculate the infection potential. Some previous studies have suggested that SARS-CoV-2 IgG antibodies remain detectable 6–9 months following infection (Figueiredo-Campos et al. 2020; Yao et al. 2021). However, recent studies have reported more substantial decreases in sensitivity in later months for certain SARS-CoV-2 antibody tests (Muecksch et al. 2021). If antibody test sensitivity decreases substantially over time, then the seroprevalence studies (and therefore MERMAID) likely underestimate the prevalence. Third, we assumed that seroprevalence studies provide unbiased estimates of the population prevalence of SARS-CoV-2 antibodies; this assumption may be violated due to convenience sampling.

MERMAID could be extended in several ways. First, MERMAID assumes that disease transmissions are independent across regions. This assumption is unlikely to substantially affect our analyses of COVID-19 in US states, as we expect imported infections to have little effect in large populations with large numbers of locally infected individuals (e.g., Russell et al. 2021). However, at the early stages of the outbreak in the United States, and at finer geographic resolutions, imported infections play a central role and cannot reasonably be ignored. MERMAID could be extended in such cases by modifying the infection potential Λ_{it} to include between-region terms. Second, we assume that the population size in each region is constant over time. This assumption is relevant for estimating the proportion of the population that is susceptible, and we believe it is reasonable for U.S. states. However, death and migration are of greater importance when modeling smaller populations. Third, we assumed that individuals are immune after they have been infected, and that all individuals who were never infected are susceptible. We accounted for immunity acquired from previous infections in the population by specifying an offset term in the R_t regression model. An additional offset could be used to account for immunity through vaccination. Alternately, these factors could be estimated directly as a component of R_t without including offsets. Fourth, MERMAID could be extended to incorporate data on COVID-19 deaths over time, which may be informative for the incidence if death-ascertainment is high and the infection fatality rate is stable across time within each region. Fifth, the model could be extended to allow overdispersion in daily infection counts. Sixth, MERMAID could be extended to explicitly model the impact of vaccinations on R_t via immunity in the population. To this end, it will be important to understand the characteristics of seroprevalence tests in vaccinated individuals, and to jointly model vaccination and natural immunity in the population. Our analysis was restricted to the U.S. data in 2020, before vaccines became available; however, such an extension may be valuable for analyzing more recent time periods, where vaccination rates are expected to be a key determinant of R_t.

Available datasets on COVID-19 reported infections, tests, and seroprevalence over time in the United States have several limitations that we did not attempt to fully address. First, daily reported infection and test counts show substantial discordance between different data sources in several states (Supplementary material S3–4). Here, we used a heuristic approach to aggregate the data across different sources and remove outliers to arrive at a consensus dataset. This highlights the importance of accurate, consistent, and coordinated data collection efforts in epidemics. Second, reported COVID-19 infection and test data show some consistent irregularities across data sources, which likely do not reflect true patterns in infections. For example, sporadic lapses in reporting and strong weekday effects are evident for many U.S. states (Supplementary material S3–4). Third, while serological surveys have provided a vital secondary means of estimating the prevalence of COVID-19, their sample sizes are relatively limited, and they may suffer from biases due to non-random convenience sampling. Seroprevalence estimates in several U.S. states are inconsistent with confirmed infection numbers, or decrease over time. For example, the August 26 – September 10, 2020 round of serological survey in South Dakota estimated 6,050 infections, which is smaller than the cumulative reported infection counts until August 26 which was 11,851 (based on COVIDTracking 2021). In New York, the seroprevalence estimates steadily decreased from 23.3% to 17% from July 31–August 11, 2020 survey round to September 11–September 24, 2020 survey round. These apparently inaccurate seroprevalence estimates may lead to unreliable inferences in MERMAID (shown for all states in Supplementary material S5).

In summary, our work provides an integrated framework to model infectious disease dynamics across one or multiple regions over time, accounting for reporting lags and under-ascertainment, and using flexible regression models for R_t. Applied to the 2020 COVID-19 data from the United States, MERMAID suggests that integrated analysis of seroprevalence and PCR test data can provide more accurate estimates of the prevalence over time. Our analysis also shows substantial reductions in COVID-19 reproductive rates associated with containment policies across the United States, and highlights difficulties disentangling the causal effects of policy from the observed data.

Software

The proposed procedure was implemented in the R package *MERMAID*, which is publicly available at *https://github.com/lin-lab/MERMAID*

Supplementary Materials

The online supplementary materials provide technical proofs and additional data analysis results.

- S1: Supplementary text including technical proofs, additional simulation study under a more realistic ascertainment scenario, and details on preprocessing and QC procedures.
- S2: Figures and tables from simulation studies.
- S3: Figures on confirmed COVID-19 infection time series.
- S4: Figures on PCR test time series.
- S5: Figures from MERMAID analysis of all US states.

Funding

This work was supported by a grant from Partners in Health. The authors thank the reviewers for helpful comments and suggestions.

ORCID

Xihong Lin http://orcid.org/0000-0001-7067-7752

References

Adam, D. (2020), "Special Report: The Simulations Driving the World's Response to COVID-19," *Nature*, 580, 316–319. [1574]

Alene, M., Yismaw, L., Assemie, M. A., Ketema, D. B., Mengist, B., Kassie, B., and Birhan, T. Y. (2021), "Magnitude of Asymptomatic COVID-19 Cases Throughout the Course of Infection: A Systematic Review and Meta-Analysis," *PloS One*, 16, e0249090. [1561]

Anderson, R. M., and May, R. M. (1992), *Infectious Diseases of Humans: Dynamics and Control*, Oxford: Oxford University Press. [1562]

Bajema, K. L., Wiegand, R. E., Cuffe, K., Patel, S. V., Iachan, R., Lim, T., Lee, A., Moyse, D., Havers, F. P., Harding, L., Fry, A. M., Hall, A. J., Martin, K., Biel, M., Deng, Y., Meyer, W. A., Mathur, M., Kyle, T., Gundlapalli, A. V., Thornburg, N. J., Petersen, L. R., and Edens, C. (2020), "Estimated SARS-CoV-2 Seroprevalence in the us as of September 2020," *JAMA Internal Medicine*, 181, 450–460. [1561,1562,1570]

Bergman, A., Sella, Y., Agre, P., and Casadevall, A. (2020), "Oscillations in US COVID-19 Incidence and Mortality Data Reflect Diagnostic and Reporting Factors," *Msystems*, 5, e00544–20. [1561]

Bertozzi, A. L., Franco, E., Mohler, G., Short, M. B., and Sledge, D. (2020), "The Challenges of Modeling and Forecasting the Spread of COVID-19. *Proceedings of the National Academy of Sciences—PNAS*, 117, 16732–16738. [1561]

Bettencourt, L. M. A., and Ribeiro, R. M. (2008), "Real time Bayesian Estimation of the Epidemic Potential of Emerging Infectious Diseases," *PloS One*, 3, e2185–e2185. [1562]

Brooks, J. T., Butler, J. C., and Redfield, R. R. (2020), "Universal Masking to Prevent SARS-CoV-2 Transmission–the Time is Now," *JAMA*, 324, 635–637. [1574]

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008), "Bootstrap-Based Improvements for Inference with Clustered Errors," *The Review of Economics and Statistics*, 90, 414–427. [1573]

CDC (2021), "United States COVID-19 Cases and Deaths by State Over Time | Data | Centers for Disease Control and Prevention," *Availableat:https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36*. (Accessed on 04/05/2021). [1561,1569,1570]

Chu, D. K., Akl, E. A., Duda, S., Solo, K., Yaacoub, S., Schünemann, H. J., El-harakeh, A., Bognanni, A., Lotfi, T., Loeb, M., et al. (2020). "Physical Distancing, Face Masks, and Eye Protection to Prevent Person-to-Person Transmission of SARS-CoV-2 and COVID-19: A Systematic Review and Meta-Analysis," *The Lancet*, 395, 1973–1987. [1574]

Cori, A., Ferguson, N. M., Fraser, C., and Cauchemez, S. (2013), "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics," *American Journal of Epidemiology*, 178, 1505–1512. [1562,1563,1564,1566,1570]

COVIDTracking (2021), "The Data: The COVID Tracking Project." Available at: *https://covidtracking.com/data* [1561,1569,1570,1575]

Currie, C. S., Fowler, J. W., Kotiadis, K., Monks, T., Onggo, B. S., Robertson, D. A., and Tako, A. A. (2020), "How Simulation Modelling can Help Reduce the Impact of COVID-19," *Journal of Simulation*, 14, 83–97. [1574]

Figueiredo-Campos, P., Blankenhaus, B., Mota, C., Gomes, A., Serrano, M., Ariotti, S., Costa, C., Nunes-Cabaço, H., Mendes, A. M., Gaspar, P., Pereira-Santos, M. C., Rodrigues, F., Condeço, J., Escoval, M. A., Santos, M., Ramirez, M., Melo-Cristino, J., Simas, J. P., Vasconcelos, E., Afonso, A. Â, Veldhoen, M. (2020), "Seroprevalence of Anti-SARS-CoV-2 Antibodies in COVID-19 Patients and Healthy Volunteers up to 6 Months Post Disease Onset," *European Journal of Immunology*, 50, 2025–2040. [1574]

Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., Monod, M., Ghani, A. C., Donnelly, C. A., Riley, S., Vollmer, M. A. C., Ferguson, N. M., Okell, L. C., and Bhatt, S. (2020), "Estimating the Effects of Non-Pharmaceutical Interventions on COVID-19 in Europe," *Nature*, 584, 257–261. [1562]

Fraser, C. (2007), "Estimating Individual and Household Reproduction Numbers in an Emerging Epidemic," *PloS One*, 2, e758. [1564]

Gostic, K. M., McGough, L., Baskerville, E. B., Abbott, S., Joshi, K., Tedijanto, C., Kahn, R., Niehus, R., Hay, J. A., De Salazar, P. M., Hellewell, J., Meakin, S., Munday, J. D., Bosse, N. I., Sherrat, K., Thompson, R. N., White, L. F., Huisman, J. S., Scire, J., Bonhoeffer, S., Stadler, T., Wallinga, J., Funk, S., Lipsitch, M., and Cobey, S. (2020), "Practical Considerations for Measuring the Effective Reproductive Number, Rt," *PLoS Computational Biology*, 16, e1008409–e1008409. [1561,1562,1565]

Hale, T., Angrist, N., Goldszmidt, R., Kira, B., Petherick, A., Phillips, T., Webster, S., Cameron-Blake, E., Hallas, L., Majumdar, S., Tatlow, H. (2021), A Global Panel Database of Pandemic Policies (Oxford COVID-19 Government Response Tracker)," *Nature Human Behaviour*, 1–10. [1572]

Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., and Wang, C. (2020), "Reconstruction of the Full Transmission Dynamics of COVID-19 in Wuhan," *Nature*, 584, 420–424. [1561,1562]

Havers, F. P., Reed, C., Lim, T., Montgomery, J. M., Klena, J. D., Hall, A. J., Fry, A. M., Cannon, D. L., Chiang, C.-F., Gibbons, A., Krapiunaya, I., Morales-Betoulle, M., Roguski, K., Rasheed, M. A. U., Freeman, B., Lester, S., Mills, L., Carroll, D. S., Owen, S. M., Johnson, J. A., Semenova, V., Blackmore, C., Blog, D., Chai, S. J., Dunn, A., Hand, J., Jain, S., Lindquist, S., Lynfield, R., Pritchard, S., Sokol, T., Sosa, L., Turabelidze, G., Watkins, S. M., Wiesman, J., Williams, R. W., Yendell, S., Schiffer, J., and Thornburg, N. J. (2020), "Seroprevalence of Antibodies to SARS-CoV-2 in 10 Sites in the United States, March 23-May 12, 2020," *JAMA Internal Medicine*, 180, 1576–1586. [1562]

He, X., Lau, E. H., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B. J., Li, F., Leung, G. M. (2020), "Temporal Dynamics in Viral Shedding and Transmissibility of COVID-19," *Nature Medicine*, 26, 672–675. [1561,1562]

HealthData.gov (2021), "COVID-19 Diagnostic Laboratory Testing (PCR Testing) Time Series; healthdata.gov." Available at: *https://healthdata.gov/dataset/COVID-19-Diagnostic-Laboratory-Testing-PCR-Testing/j8mb-icvb*. [1570]

Hernán, M. A., Brumback, B. A., and Robins, J. M. (2002), "Estimating the Causal Effect of Zidovudine on CD4 Count With a Marginal Structural Model for Repeated Measures," *Statistics in Medicine*, 21, 1689–1709. [1573]

Inglesby, T. V. (2020), "Public Health Measures and The Reproduction Number of SARS-CoV-2," *Journal of the American Medical Association*, 323, 2186–2187. [1561]

Iwasaki, A. (2021), "What Reinfections Mean for COVID-19," *The Lancet Infectious Diseases*, 21, 3–5. [1562]

Jewell, N. P., Lewnard, J. A., and Jewell, B. L. (2020), "Predictive Mathematical Models of the COVID-19 Pandemic: Underlying Principles and Value of Projections," *Journal of the American Medical Association*, 323, 1893–1894. [1561]

JHU-CSSE (2021), "COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University." Available at: *https://github.com/CSSEGISandData/COVID-19*. (Accessed on 04/05/2021). [1561]

Lange, K. (1995a), "A Gradient Algorithm Locally Equivalent to the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 57, 425–437. [1566]

——— (1995b), "A Quasi-Newton Acceleration of the EM Algorithm," *Statistica Sinica*, 5, 1–18. [1566]

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020), "The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application," *Annals of Internal Medicine*, 172, 577–582. [1570]

Lee, L. Y., Rozmanowski, S., Pang, M., Charlett, A., Anderson, C., Hughes, G. J., Barnard, M., Peto, L., Vipond, R., Sienkiewicz, A., Hopkins, S., Bell, J., Crook, D. W., Gent, N., Walker, A. S., Peto, T. E. A., Eyre, D. W. (2021), "SARS-CoV-2 Infectivity by Viral Load, S Gene Variants and Demographic Factors and the Utility of Lateral Flow Devices to Prevent Transmission," medRxiv. [1574]

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 44, 226–233. [1566]

Lu, F. S., Nguyen, A. T., Link, N. B., Davis, J. T., Chinazzi, M., Xiong, X., Vespignani, A., Lipsitch, M., and Santillana, M. (2020), "Estimating the Cumulative Incidence of COVID-19 in the United States Using Four Complementary Approaches," medRxiv. [1562]

Manski, C. F., and Molinari, F. (2021), "Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem," *Journal of Econometrics*, 220, 181–192. [1562]

Miller, A. C., Hannah, L., Futoma, J., Foti, N. J., Fox, E. B., D'Amour, A., Sandler, M., Saurous, R. A., and Lewnard, J. A. (2020), "Statistical Deconvolution for Inference of Infection Time Series," medRxiv. [1562,1565]

Muecksch, F., Wise, H., Batchelor, B., Squires, M., Semple, E., Richardson, C., McGuire, J., Clearly, S., Furrie, E., Greig, N., et al. (2021), "Longitudinal Serological Analysis and Neutralizing Antibody Levels in Coronavirus Disease 2019 Convalescent Patients," *The Journal of Infectious Diseases*, 223, 389–398. [1574]

Murray, C. J., and Piot, P. (2021), "The Potential Future of the COVID-19 Pandemic: Will SARS-CoV-2 Become a Recurrent Seasonal Infection?" *JAMA*, 325, 1249–1250. [1562]

Ndaïrou, F., Area, I., Nieto, J. J., and Torres, D. F. (2020), "Mathematical Modeling of COVID-19 Transmission Dynamics With a Case Study of Wuhan," *Chaos, Solitons and Fractals*, 135, 109846–109846. [1562]

Nishiura, H., Linton, N. M., and Akhmetzhanov, A. R. (2020), "Serial Interval of Novel Coronavirus (COVID-19) Infections," *International Journal of Infectious Diseases*, 93, 284–286. [1567,1570]

Oakes, D. (1999), "Direct Calculation of the Information Matrix Via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 61, 479–482. [1566]

Pan, A., Liu, L., Wang, C., Guo, H., Hao, X., Wang, Q., Huang, J., He, N., Yu, H., Lin, X., Wei, S., and Wu, T. (2020), "Association of Public Health Interventions With the Epidemiology of the COVID-19 Outbreak in Wuhan, China," *Journal of the American Medical Association*, 323, 1915–1923. [1562]

Pei, S., Kandula, S., and Shaman, J. (2020), "Differential Effects of Intervention Timing on COVID-19 Spread in the United States," *Science Advances*, 6, eabd6370. [1562]

Petermann, M., and Wyler, D. (2020), "A Pitfall in Estimating the Effective Reproductive Number Rt for COVID-19," *Swiss Medical Weekly*, 150. [1562,1565,1566,1567]

Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2009), "Fully Exponential Laplace Approximations for the Joint Modelling of Survival and Longitudinal Data," *Journal of the Royal Statistical Society*, 71, 637–654. [1566]

Roda, W. C., Varughese, M. B., Han, D., and Li, M. Y. (2020), "Why Is It Difficult to Accurately Predict the COVID-19 Epidemic?" *Infectious Disease Modelling*, 5, 271–281. [1561]

Rojas, D. P., Dean, N. E., Yang, Y., Kenah, E., Quintero, J., Tomasi, S., Ramirez, E. L., Kelly, Y., Castro, C., Carrasquilla, G., Halloran, M. E., Longini, I. M. (2016), "The Epidemiology and Transmissibility of Zika Virus in Girardot and San Andres Island, Colombia, September 2015 to January 2016," *Eurosurveillance*, 21, 30283. [1564]

Russell, T. W., Wu, J. T., Clifford, S., Edmunds, W. J., Kucharski, A. J., Jit, M., et al. (2021), "Effect of Internationally Imported Cases on Internal Spread of COVID-19: A Mathematical Modelling Study," *The Lancet Public Health*, 6, e12–e20. [1574]

Schechtman, K. W. (2021), "Analysis & Updates Federal Testing Data's Last Mile." Available at: *https://covidtracking.com/analysis-updates/federal-testing-datas-last-mile*. [1561,1562,1570]

Steele, B. M. (1996), "A Modified EM Algorithm for Estimation in Generalized Mixed Models," *Biometrics*, 52, 1295–1310. [1566]

Tian, T., Tan, J., Jiang, Y., Wang, X., and Zhang, H. (2021), "Evaluate the Risk of Resumption of Business for the States of New York, New Jersey and Connecticut Via a Pre-Symptomatic and Asymptomatic Transmission Model of COVID-19," *Journal of Data Science*, 19, 178–196. [1562]

Tierney, L., Kass, R. E., and Kadane, J. B. (1989), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716. [1566]

USAFacts (2021), "US COVID-19 Cases and Deaths by State." Available at *https://usafacts.org/visualizations/coronavirus-covid-19-spread-map*. [1561,1569,1570]

van Kampen, J. J., van de Vijver, D. A., Fraaij, P. L., Haagmans, B. L., Lamers, M. M., Okba, N., van den Akker, J. P., Endeman, H., Gommers, D. A., Cornelissen, J. J., Hoek, R. A. S., van der Eerden, M. M., Hesselink, D. A., Metselaar, H. J., Verbon, A., de Steenwinkel, J. E. M., Aron, G. I., van Gorp, E. C. M., van Boheemen, S., Voermans, J. C., Boucher, C. A. B., Molenkamp, R., Koopmans, M. P. G., Geurtsvankessel, C., van der Eijk, A. A.(2021), "Duration and Key Determinants of Infectious Virus Shedding in Hospitalized Patients With Coronavirus Disease-2019 (COVID-19)," *Nature Communications*, 12, 1–6. [1562]

Wallinga, J., and Teunis, P. (2004), "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures," *American Journal of Epidemiology*, 160, 509–516. [1562,1564]

Wölfel, R., Corman, V. M., Guggemos, W., Seilmaier, M., Zange, S., Müller, M. A., Niemeyer, D., Jones, T. C., Vollmar, P., Rothe, C., Hoelscher, M., Bleicker, T., Brünink, S., Schneider, J., Ehmann, R., Zwirglmaier, K., Drosten, C., Wendtner, C. (2020), "Virological Assessment of Hospitalized Patients With COVID-2019," *Nature*, 581, 465–469. [1562]

Yao, L., Wang, G. L., Shen, Y., Wang, Z. Y., Zhan, B. D., Duan, L. J., Lu, B., Shi, C., Gao, Y. M., Peng, H. H., Wang, G. Q., Wang, D. M., Jiang, M. D., Cao, G. P., and Ma, M. J. (2021), "Persistence of Antibody and Cellular Immune Responses in COVID-19 Patients Over Nine Months After Infection," *The Journal of Infectious Diseases*, 224, 586-594. [1574]