Health Care Management Science (2021) 24:375–401
https://doi.org/10.1007/s10729-021-09555-3

The development and deployment of a model for hospital-level COVID-19 associated patient demand intervals from consistent estimators (DICE)

Linying Yang[1] · Teng Zhang[2] · Peter Glynn[2] · David Scheinker[2]

Received: 6 November 2020 / Accepted: 3 February 2021 / Published online: 22 March 2021

Abstract

Hospitals commonly project demand for their services by combining their historical share of regional demand with forecasts of total regional demand. Hospital-specific forecasts of demand that provide prediction intervals, rather than point estimates, may facilitate better managerial decisions, especially when demand overage and underage are associated with high, asymmetric costs. Regional point forecasts of patient demand are commonly available, e.g., for the number of people requiring hospitalization due to an epidemic such as COVID-19. However, even in this common setting, no probabilistic, consistent, computationally tractable forecast is available for the fraction of patients in a region that a particular institution should expect. We introduce such a forecast, DICE (Demand Intervals from Consistent Estimators). We describe its development and deployment at an academic medical center in California during the 'second wave' of COVID-19 in the Unite States. We show that DICE is consistent under mild assumptions and suitable for use with perfect, biased and unbiased regional forecasts. We evaluate its performance on empirical data from a large academic medical center as well as on synthetic data.

Keywords COVID-19 · Hospital-level forecast · Prediction interval · Parametric bootstrap · Moment method · Prediction bias

Highlights

- Hospital managers require forecasts of the number of people requiring hospitalization for COVID-19 at their institution, but such forecasts are available only at the level of county or state.
- DICE is a probabilistic model that converts regional estimates into hospital-specific forecasts.
- DICE provides point forecasts along with prediction intervals that incorporate uncertainty about the accuracy of the regional forecast and uncertainty about the fraction of the patients in the region that will go to a particular hospital.

✉ Linying Yang
yanglinying1024@gmail.com

[1] Institute for Computational and Mathematical Engineering, Stanford University, Stanford, CA 94305, USA

[2] Department of Management Science and Engineering, Stanford University, Stanford, CA 94305, USA

1 Introduction

The COVID-19 pandemic has disrupted hospital operations the world over. Large influxes of patients requiring intensive care and mechanical ventilation have overwhelmed capacity, forced hospitals to triage, and have been associated with significantly elevated case fatality rates. Shortages of personal protective equipment (PPE) have exposed healthcare workers to additional risk and many have contracted COVID-19 and died.

Hospitals managers have a variety of options to increase total and available capacity when planning for an influx of COVID-19 patients [1]. Managers may be able to increase total capacity by calling in additional nurses and doctors, opening previously closed beds, and acquiring additional PPE. Managers may be able to increase available capacity by expediting patient discharge or canceling or delaying non-discretionary, non-urgent patient admissions [2]. The potential detriments to the quality of care and higher costs associated with these actions may be partially or fully mitigated if the decision to act is made with sufficient lead time. In the worst-case scenario when a hospital has

insufficient intensive care unit (ICU) or ventilator capacity, patients with COVID-19 may experience significantly higher case mortality rates [3]. In less dire scenarios, nurses called in to work on short notice may require overtime pay while those scheduled a week in advance may not; PPE is less expensive when its purchase is not expedited; and patients whose non-urgent procedures are scheduled for later will experience less disruption than patients whose procedures are cancelled on short notice. In the United States, where healthcare is paid for through a combination of private and public insurance, the pandemic has created the additional challenge of significant financial stress as COVID-19 patients are associated with lower rates of reimbursement than patients who receive non-urgent, non-discretionary procedures such as tumor removal surgery or chemotherapy [4].

The complementary challenges of ensuring sufficient capacity to meet the demand associated with COVID-19 while avoiding unnecessarily long delays to non-COVID-19 care, require hospital managers to generate forecasts of the volume of COVID-19 patients requiring care at their institution. Managerial decisions based on forecasts of COVID-19 may benefit from the availability of the forecast with as much lead time as possible. To allow managers to account for the asymmetric risk associated with having insufficient capacity to meet urgent COVID-19 demand or non-urgent procedural demand, such forecasts should provide probabilistic, rather than point, estimates.

Our methodology reflects the random fluctuations that arise at the hospital level that are averaged out at the regional level. For example, if a hospital receives, on average, 5% of a county's hospitalization and the forecast county hospitalization level is 100, the random fluctuation about the mean hospital load of 5 patients can be significant (in a relative error sense). In particular, our methodology provides a prediction interval on the number of COVID-19 positive patients at a given hospital rather than a "point forecast". In addition, our methodology takes into account the additional uncertainty induced by estimation error associated with estimating the underlying statistical parameters from observed data.

This paper is concerned with developing statistical methods to support hospital decision making with regard to COVID-19 capacity planning issues. In particular, hospital leadership can benefit from statistical tools to help them assess the amount of capacity that will need to be assigned to coronavirus patients in the weeks to come. A serious complication is that epidemiological forecasts typically focus on aggregate COVID-19 predictions that are provided at the regional level. For example, in California, the available COVID-19 forecasts are provided at the county level. Our goal in this paper is to provide a statistically principled methodology for obtaining hospital-level coronavirus hospitalization forecasts from such regional forecasts. Such forecasts are more useful than regional ones, for example, for manager preparing for an influx of COVID-19 patients to a busy hospital that has capacity available to simultaneously accommodate up to 20 COVID-19 patients, would have to call in further staff to accommodate 21-30 COVID-19 patients, and would have to call in additional staff and cancel scheduled procedures to accommodate over 30 COVID-19 patients.

Given a regional forecast for the daily number of hospitalizations as well as historical data on the share of regional hospitalizations accommodated by a specific hospital, all assumed to be Poisson random variables, we develop a forecast model DICE (Demand Intervals from Consistent Estimators). The model intentionally is "lightweight" in terms of the data needed to make predictions: only county level forecasts and actual hospitalizations, plus local hospital-level hospitalization numbers. We take the view that the epidemiology community is best suited to model county-level hospitalizations. Such forecasts take into account local measures to reduce contacts, county-level age distribution, the number of patients testing positive, etc. One challenge in producing prediction intervals in this setting is that the quantity being predicted is count-level data that is integer-valued. Especially when the number of hospitalizations is small, this integrality plays a central role in generating good prediction intervals. We note that the SIR models that are widely used produce point forecasts that are non-integer. This requires the building of a principled approach to convert forecasts based on continuous modeling methods into a prediction interval for a stochastic integer-valued quantity. Another big issue is that the method needs to deal with an underlying phenomenon that has dynamics that can exhibit periods of quiescence, exponential growth, and gradual decay, so does not exhibit the stationarity that is generally assumed in the literature.

The primary contributions of this paper are as follows.

- We show that DICE is consistent under mild assumptions and suitable for use with biased and unbiased regional forecasts.
- We show that DICE performed well on empirical data from a large academic medical center in California as well as on synthetic data.
- We describe the COVID-19 related capacity management decisions facilitated by the use of DICE.

The rest of the paper is organized as follows. Section 2 reviews the related literature. Section 3 outlines the model setting. Sections 4 to 7 describe the methods of generating prediction intervals under three assumptions about the county-level predictors: perfect forecast, unbiased forecast and biased forecast. Section 8 reports the empirical findings. Further discussions and conclusions can be found in Section 9.

2 Literature

Numerous COVID-19 forecasting models have been developed since the start of the pandemic. A lot of them forecast regional-level COVID-19 cases, hospitalizations and deaths [5–12] and [13].[1] Most such models use publicly available data and epidemic models to forecast hospitalizations down to the level of a single *county* or several adjacent counties. However, few tools are available for *hospitals* to make a probabilistic forecast of their expected share of the forecast regional volume. The data available to make such a forecast include: the outputs of the aforementioned models; detailed historical data on county-level hospitalizations, available from the national authorities such as [14]; real-time data on hospitalizations in a particular county or region available from local authorities such as [15]; and hospital-specific hospitalization data available to the managers of the institution generating the forecast.

Work on epidemic/influenza forecasting has examined national/state level [16, 17] and regional level [18–20] forecasts. The most relevant research on the *hospital* level we could find are [21, 22] and [23], where the authors use historical data and public available data to generate hospital influenza visits. Our work complements the prior work in several ways: 1. most papers only generate point estimates, while we provide prediction intervals; 2. apart from historical data, these studies use numerous sources of public data including the Google influenza index and Twitter posts, while we require only projections generated by regional forecasts; 3. since our model can make use of any forecast, there is no additional effort necessary to compare performance based on several forecasts. To our best knowledge, this is the first model to generate integral hospital-level forecasts with prediction intervals based on regional projections.

From the broader literature on time series forecasting, we summarize how the model presented differs from several existing classes of methods:

1. Classic auto-regressive models [24, 25]. These models assume a linear auto-regressive relationship with random noise. However, such models are not well suited for nonlinear and non-stationary processes such as the spread of COVID and do not incorporate information from outside the time series such as an external forecasts. Also, they model continuous random quantities, not integer value quantities.
2. Hidden Markov models (HHMs) [26, 27]. This is a special class of mixture models, where the observed time series is structured as a function of the underlying, unobserved states. However, it is usually computational expensive to estimate such models and HMMs may perform poorly in non-stationary settings such as COVID [28].
3. Neural networks [29, 30]. This is a class of nonlinear parametric time series forecasting models that are applied to areas including finance, energy, and manufacturing. Far more data are typically required to fit neural nets than what is available or used in our setting. Further, such models generally do not produce prediction intervals.
4. Susceptible exposed infected recovered (SEIR) models [31, 32]. Such methods explicitly model the dynamics of the data generating process using differential equations. SEIR models are primarily designed for large populations rather than individual institutions. Also, most widely used SEIR models are generally deterministic, not stochastic, and they produce point forecasts that are non-integer. This requires the building of a principled approach towards converting the forecasts based on continuous modeling methods into a prediction interval for a stochastic integer-valued quantity. We decide to use such models as our underlying forecast models.
5. Discrete Event Simulation (DES) [33–35]. DES is a technique that has been used to model the flow of patients into a hospital based on historical patient data and detailed, hospital- and region-specific assumptions about resource consumption. Such methods require numerous ad hoc, rather than principled, modeling choices and are of limited generalizeability beyond the setting in which they are designed while we provide a more principled, widely applicable approach. Also, these methods only measure "interval stochasticity"; they do not compute calibration error, for example. Our model also considers calibration error.

3 Setting

This model was developed in response to a request from the COVID-19 planning leadership of a large academic medical center (AMC) in a large county in California during the summer of 2020. After the initial wave of COVID-19 cases was brought under control with non-pharmaceutical interventions such as social distancing, the hospital restarted non-urgent admissions for procedures such as surgery. As national news of a "second wave" of COVID-19 hospitalizations spread, the AMC leadership wanted to prepare. They requested a forecast that would inform them, with as much notice as possible, of an influx of COVID-19 patients sufficiently large that elective admissions should be halted in order to make capacity available for the expected

[1] More models can be found on the CDC website https://www.cdc.gov/coronavirus/2019-ncov/covid-data/mathematical-modeling.html, where it reports forecasts from 32 models on *case* forecasts, 47 models on *death* forecasts and 12 models on *hospitalization* forecast.

COVID-19 patients. We were provided with the hospital's historical data on the number of admissions and the length of stay of each patient in the ACU and ICU, historical and forecast data for the total number of hospitalizations in the county, and automated daily updates on the number of new COVID-19 admissions to the ICU and ACU as well as patients currently in those units. We worked with hospital leadership to estimate the capacity of COVID-19 patients that the institution could accommodate without having to increase available capacity by canceling scheduled procedures. We also worked with the leadership to determine an order for cancelling scheduled, non-urgent surgical procedures if necessary. The order was based primarily on the clinical acuity of those requiring the procedure, the average ICU and ACU post-operative length of stay associated with the procedure, and additional constraints on hospital operations. The specifics of the hospital operational planning efforts are likely to vary significantly across institutions and are outside the scope of this work.

The goal of the present work was to generate a forecast of patient demand based on recent data on the share of all COVID-19 patients in the county. One specific use of the forecast would be to provide two weeks notice that the institution may have to cancel scheduled procedures in order to accommodate the demand for beds by COVID-19 patients. Since hospital occupancy fluctuates naturally, rather than determine a hard cut-off for cancelling procedures, hospital leadership requested that we notify them if the upper bound of the prediction interval exceeded a pre-specified lower bound at which point they would evaluate the prospect of cancelling cases.

4 Prediction intervals with perfect forecasts

We start by describing the problem setting from a mathematical perspective. We assume that we are currently in day 0 and have been tasked with producing prediction intervals for the future number of hospital-level acute care unit (ACU) and intensive care unit (ICU) hospitalization at the end of day r, with $r \geq 0$. For the purpose of predicting these hospital-level prediction intervals, we have available historical data $((A_j, B_j, N_j) : -n \leq j \leq -1)$, where N_j is the total number of regional hospitalizations at the end of day j, A_j is the number of acute care hospitalizations at the given hospital at the conclusion of day j, and B_j is the number of ICU hospitalizations at the given hospital at the end of day j. Furthermore, we assume that we have available a point forecast F_r for the mean number of regional hospitalizations at the end of day r.

Throughout the paper, we take the view that the A_j's,B_j's and N_j's can be reasonably modeled as Poisson distributed random variables (rv's).We will use the notation $\mathcal{P}(\lambda)$ to denote a Poisson rv with mean λ. There is an extensive mathematical theory supporting the use of Poisson rv's in the setting of such count statistics; see, for example, [36].

A simple model relates A_j and B_j to N_j by assuming that $\mathbb{E}A_j = p_0\mathbb{E}N_j$ and $\mathbb{E}B_j = q_0\mathbb{E}N_j$ for $p_0, q_0 \geq 0$. Because the N_j's are subject to episodic epidemic growth spurts, *we do not assume that* $\mathbb{E}N_j$ *is constant.* Instead, we permit $\lambda_j \equiv \mathbb{E}N_j$ to fluctuate in a potentially complex fashion.

In this section, we assume that point forecast F_r is *perfect*, in the sense that

$$F_r = \lambda_r. \tag{4.1}$$

It follows that if we select $l(\lambda)$ (the lower endpoint) as the largest integer such that $P(\mathcal{P}(\lambda) < l(\lambda)) \leq \frac{\delta}{2}$ and $u(\lambda)$ (the upper endpoint) as the smallest integer such that $P(\mathcal{P}(\lambda) > u(\lambda)) \leq \frac{\delta}{2}$, then $[l(F_r), u(F_r)]$ is a $100(1-\delta)\%$ *prediction interval* for N_r having the property that $P(N_r \in [l(F_r), u(F_r)]) \geq 1 - \delta$.

To obtain similar prediction intervals for A_r and B_r, we need to estimate p_0 and q_0 from the data. The obvious estimators for p_0 and q_0 are given by

$$\begin{aligned} \hat{p} &= \sum_{j=-n}^{-1} A_j / \sum_{j=-n}^{-1} N_j, \\ \hat{q} &= \sum_{j=-n}^{-1} B_j / \sum_{j=-n}^{-1} N_j. \end{aligned} \tag{4.2}$$

In fact, $\hat{p}$ and $\hat{q}$ are the maximum likelihood estimators (MLE's) for p_0 and q_0 when A_j (and B_j) are, conditional on $N_{-n}, \ldots, N_{-1}$, independent in j and binomially distributed with parameters N_j and p_0 (and q_0).

This leads to the prediction intervals $[l(\hat{p}F_r), u(\hat{p}F_r)]$ for A_r and $[l(\hat{q}F_r), u(\hat{q}F_r)]$ for B_r. We refer to these prediction intervals as the *plug-in prediction intervals* based on perfect forecasts.

5 Prediction intervals with perfect forecasts: incorporating estimation uncertainty

Our frequentist approach starts by setting $\theta = (p, q)$ and letting $P_\theta(\cdot)$ be the probability model under which the $(A_i, B_i, N_i - A_i - B_i)$'s are conditionally independent given the N_j's, with $(A_i, B_i, N_i - A_i - B_i)$ following a multinomial distribution with parameters $(N_i, p_0, q_0, 1 - p_0 - q_0)$. Our ideal prediction interval for A_r would, of course, be the interval $[\ell(p_0F_r), u(p_0F_r)]$. Since p_0 is unknown, the plug-in interval $[\ell(\hat{p}F_r), u(\hat{p}F_r)]$ of

Section 4 is an obvious alternative. However, because $\hat{p}$ is random, we can not guarantee that

$$P_{\theta_0}(A_r < \ell(\hat{p}F_r)) \leq \delta/2, \tag{5.1}$$

where $\theta_0 = (p_0, q_0)$.

Instead, we seek a probabilistic guarantee, namely that Eq. 5.1 holds, with probability (or confidence level) $1-\alpha$.

We can accomplish this by choosing the integer z_ℓ so that

$$P_{\theta_0}\left(\ell(\hat{p}F_r) - \ell(p_0F_r) \leq z_\ell\right) \geq 1-\alpha. \tag{5.2}$$

On the event $\{\ell(p_0F_r) \geq \ell(\hat{p}F_r) - z_\ell\}$,

$$P_{\theta_0}\left(A_r < \ell(\hat{p}F_r) - z_\ell\right) \leq P_{\theta_0}\left(A_r < \ell(p_0F_r)\right) \leq \delta/2.$$

Hence, with confidence at least $1-\alpha$, $\ell(\hat{p}F_r) - z_\ell$ is an appropriately chosen value for the left endpoint of A_r's prediction interval.

Similarly, if we choose the integer z_r so that

$$P_{\theta_0}\left(u(\hat{p}F_r) - u(p_0F_r) \geq z_r\right) \geq 1-\alpha, \tag{5.3}$$

$u(\hat{p}F_r) - z_r$ is a right endpoint for which $P_{\theta_0}\left(A_r > u(\hat{p}F_r) - z_r\right) \leq \delta/2$ holds, at a confidence level of at least $1-\alpha$. Hence, we adopt the interval $[\ell(\hat{p}F_r) - z_\ell, u(\hat{p}F_r) - z_r]$ as our prediction interval for A_r that takes into account the estimation uncertainty that is present in $\hat{p}$.

To compute z_ℓ and z_r from Eqs. 5.2 and 5.3, we use the parametric bootstrap (see, for example, [37]), thereby computing the values z_ℓ^* and z_r^* such that

$$P_{\hat{\theta}}\left(\ell(\hat{p}^*F_r) - \ell(\hat{p}F_r) \leq z_\ell^*\right) \geq 1-\alpha$$

and

$$P_{\hat{\theta}}\left(u(\hat{p}^*F_r) - u(\hat{p}F_r) \geq z_r^*\right) \geq 1-\alpha,$$

where $\hat{\theta} = (\hat{p}, \hat{q})$ and $\hat{p}^*$ is the estimator for $\hat{p}$ obtained from a bootstrap sample of the data set; the details can be found in the algorithm as described below. This leads to the prediction interval $[\ell(\hat{p}F_r) - z_\ell^*, u(\hat{p}F_r) - z_r^*]$; we refer to this as the *bootstrap prediction interval* for A_r based on perfect forecasts. We can similarly compute the bootstrap prediction interval for B_r based on perfect predictions.

Specifically, our bootstrap prediction intervals are produced by the following algorithm.

Algorithm 1

1. Simulate independent Poisson random variables $(N_i^* : -n \leq i \leq -1)$ with mean $(F_i : -n \leq i \leq -1)$.
2. Conditional on N_i^*, simulate a multinomial rv $(A_i^*, B_i^*, N_i^* - A_i^* - B_i^*)$ with parameters $(n, \hat{p}, \hat{q}, 1-\hat{p}-\hat{q})$.
3. Compute

$$\hat{p}^* = \frac{\sum_{j=-n}^{-1} A_j^*}{\sum_{j=-n}^{-1} N_j^*},$$

$$\hat{q}^* = \frac{\sum_{j=-n}^{-1} B_j^*}{\sum_{j=-n}^{-1} N_j^*}.$$

4. Compute $(\ell(\hat{p}^*F_r), u(\hat{p}^*F_r), \ell(\hat{q}^*F_r), u(\hat{q}^*F_r))$.
5. Repeat steps 1 to 4 b times, thereby yielding b 4-tuples $(\ell(\hat{p}_i^*F_r), u(\hat{p}_i^*F_r), \ell(\hat{q}_i^*F_r), u(\hat{q}_i^*F_r))$.
6. Compute the smallest integers $z_{A,\ell}^*$ and $z_{B,\ell}^*$ for which

$$\frac{1}{b}\sum_{i=1}^{b} I\left(\ell(\hat{p}_i^*F_r) - \ell(\hat{p}F_r) \leq z_{A,\ell}^*\right) \geq 1-\alpha$$

and

$$\frac{1}{b}\sum_{i=1}^{b} I\left(\ell(\hat{q}_i^*F_r) - \ell(\hat{q}F_r) \leq z_{B,\ell}^*\right) \geq 1-\alpha,$$

and the largest integers $z_{A,r}^*$ and $z_{B,r}^*$ for which

$$\frac{1}{b}\sum_{i=1}^{b} I\left(u(\hat{p}_i^*F_r) - u(\hat{p}F_r) \geq z_{A,r}^*\right) \geq 1-\alpha$$

and

$$\frac{1}{b}\sum_{i=1}^{b} I\left(u(\hat{q}_i^*F_r) - u(\hat{q}F_r) \geq z_{B,r}^*\right) \geq 1-\alpha.$$

Then, the intervals $[[\ell(\hat{p}F_r) - z_{A,\ell}^*]^+, u(\hat{p}F_r) - z_{A,r}^*]$ and $[[\ell(\hat{q}F_r) - z_{B,\ell}^*]^+, u(\hat{q}F_r) - z_{B,r}^*]$ are the bootstrap prediction intervals for A_r and B_r respectively, where $[x]^+ \triangleq \max(x, 0)$ for $x \in \mathbb{R}$.

6 Unbiased forecasts with lognormal errors

The model described in Sections 4 and 5 assumes no forecast error. As a consequence, the distribution for N_i is Poisson distributed with mean F_i. However, the forecast F_i itself is imperfect, and there typically is additional uncertainty in the prediction of N_i (beyond the stochastic variability of a Poisson rv) that should be reflected in the prediction interval. In this section, we model the forecast error by assuming that

$$F_i = \lambda_i \Gamma_i^{-1}, \tag{6.1}$$

where N_i is (again) Poisson with mean λ_i, and the relative forecast error Γ_i^{-1} is assumed to be log-normally

distributed. Furthermore, we assume that the N_i's are independent of the Γ_i^{-1}'s, and that the forecasts are *relatively unbiased*, in the sense that

$$\mathbb{E}\left(\frac{N_i}{F_i}\right) = 1 \tag{6.2}$$

for all i, thereby implying that $\mathbb{E}[\Gamma_i] = 1$.

Of course, one expects that if the forecast under-predicts N_i at time i, F_{i+1} is also likely to under-predict N_{i+1}. This suggests that the Γ_i's should be modeled as a correlated sequence. In particular, we will assume that if $Y_i = \log \Gamma_i$, the Y_i's form a stationary sequence that evolves according to the recursion

$$Y_{i+1} = \rho_0 Y_i + Z_{i+1},$$

where the Z_i's are independent and identically distributed (iid) normally distributed rv's with mean μ_0 and variance σ_0^2. Note that the stationarity of the Y_i's implies that $\rho \in (-1, 1)$, with Y_i having a normal distribution having mean $\mu_0(1-\rho_0)^{-1}$ and variance $\sigma_0^2(1-\rho_0^2)^{-1}$; see [38].

For this model, we need to estimate the parameters μ_0, σ_0^2 and ρ_0 associated with the log-normally distributed forecast error sequence. As in Sections 4 and 5, we assume that we have observed the time series $((A_i, B_i, A_i, F_i) : -n \le i \le -1)$, and we adopt the view that we wish to impose as few assumptions as possible on the λ_i's (given the episodic nature of the coronavirus epidemic). For this reason, we will use the method of moments to estimate μ_0, σ_0^2 and ρ_0.

Given Eq. 6.2, we require that

$$\begin{aligned}\mathbb{E}[\exp(Y_i)] &= \exp\left(\tfrac{\mu_0}{1-\rho_0} + \tfrac{1}{2}\tfrac{\sigma_0^2}{1-\rho_0^2}\right)\\ &\triangleq m_1(\mu_0, \sigma_0^2, \rho_0) = 1.\end{aligned}$$

To obtain a second equation, note that

$$\begin{aligned}\mathbb{E}\left[\frac{N_i^2 - N_i}{F_i^2}\right] &= E[N_i^2 - N_i] \cdot \lambda_i^{-2} E[\Gamma_i^2]\\ &= ((\lambda_i + \lambda_i^2) - \lambda_i) \cdot \lambda_i^{-2} E[\Gamma_i^2]\\ &= E[\exp(2Y_i)]\\ &= \exp\left(\frac{2\mu_0}{1-\rho_0} + 2\frac{\sigma_0^2}{1-\rho_0^2}\right)\\ &\triangleq m_2(\mu_0, \sigma_0^2, \rho_0).\end{aligned}$$

For the third equation, we observe that

$$\begin{aligned}\mathbb{E}\left[\frac{N_i N_{i+1}}{F_i F_{i+1}}\right] &= \mathbb{E}N_i \cdot \mathbb{E}N_{i+1} \cdot (\lambda_i \lambda_{i+1})^{-1} \mathbb{E}\Gamma_i \Gamma_{i+1}\\ &= \mathbb{E}[\exp(Y_i + Y_{i+1})]\\ &= \mathbb{E}[\exp((1+\rho_0)Y_i + Z_{i+1})]\\ &= \exp\left(\frac{2\mu_0}{1-\rho_0} + \frac{\sigma_0^2}{1-\rho_0}\right)\\ &\triangleq m_3(\mu_0, \sigma_0^2, \rho_0).\end{aligned}$$

This suggests that we estimate μ_0, σ_0^2 and ρ_0 by minimizing the objective

$$\left(\hat{M}_2 - m_2(\mu, \sigma^2, \rho)\right)^2 + \left(\hat{M}_3 - m_3(\mu, \sigma^2, \rho)\right)^2$$

subject to

$$\begin{aligned}m_1(\mu, \sigma^2, \rho) &= 1,\\ -1 \le \rho &\le 1,\\ \sigma^2 &\ge 0,\end{aligned}$$

where

$$\begin{aligned}\hat{M}_2 &= \tfrac{1}{n}\sum_{i=-n}^{-1}\left(\tfrac{N_i^2 - N_i}{F_i^2}\right),\\ \hat{M}_3 &= \tfrac{1}{n-1}\sum_{i=-n+1}^{-1}\left(\tfrac{N_i N_{i-1}}{F_i F_{i-1}}\right),\end{aligned}$$

followed by utilizing the minimizer $(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})$ as our estimator of $(\mu_0, \sigma_0^2, \rho_0)$, and then estimate $\hat{p}$ and $\hat{q}$ as in Eq. 4.2. When n is large (and the statistical model describes the data well), we expect that the objective function will vanish at $(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})$, in which case

$$m_i(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}) = \hat{M}_i$$

will be satisfied as equations for $i = 2, 3$. In the Appendix, we prove that our estimators for μ_0, σ_0^2, and ρ_0 are consistent, under very moderate assumptions on the λ_i's.

We note that in this model, the prediction interval for N_r must reflect the additional randomness stemming from the fact that the mean of the Poisson random variable is itself random, namely it is given by $F_r\Gamma_r$. In particular, let $\mathcal{P}(\mu, \sigma^2, \rho, f)$ be a rv that is conditionally Poisson distributed, with (random) mean $f\exp(N(\mu/(1-\rho), \sigma^2/(1-\rho^2)))$, where $N(\mu/(1-\rho), \sigma^2/(1-\rho^2))$ is a normal rv with mean $\mu/(1-\rho)$ and variance $\sigma^2/(1-\rho^2)$. The plug-in prediction interval for A_r based on this model is the interval $[\ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p}F_r), u(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p}F_r))]$, where $\ell(\mu, \sigma^2, \rho, f)$ is the largest integer j such that $P(\mathcal{P}(\mu, \sigma^2, \rho, f) < j) \le \delta/2$ and $u(\mu, \sigma^2, \rho, f)$ is the smallest integer k such that $P(\mathcal{P}(\mu, \sigma^2, \rho, f) > k) \le \delta/2$. Similarly, $[\ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q}F_r), u(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q}F_r))]$ is the plug-in prediction interval for B_r.

The computation of $\ell(\mu, \sigma^2, \rho, f)$ and $u(\mu, \sigma^2, \rho, f)$ can be implemented via Monte Carlo, using the following algorithm.

Algorithm 2

1. Simulate Y_r as a normal rv with mean $\mu/(1-\rho)$ and variance $\sigma^2/(1-\rho^2)$.
2. Generate N_r as a Poisson rv with mean $f\exp(Y_r)$.
3. Repeat Steps 1 and 2, independently, m times, thereby yielding $N_{r,1}, \ldots, N_{r,m}$.
4. Define the estimator $\hat{\ell}(\mu, \sigma^2, \rho, f)$ for $\ell(\mu, \sigma^2, \rho, f)$ as the largest integer j such that

$$\frac{1}{m}\sum_{i=1}^{m} I(N_{r,i} < j) \le \delta/2$$

and define the estimator $\hat{u}(\mu, \sigma^2, \rho, f)$ for $u(\mu, \sigma^2, \rho, f)$ as the smallest integer k for which

$$\frac{1}{m}\sum_{i=1}^{m} I(N_{r,i} > k) \le \delta/2.$$

We now turn to the construction of prediction intervals for A_r and B_r that reflect the additional uncertainty due to the need to estimate $\mu_0, \sigma_0^2, \rho_0$ fro the observed data $((A_i, B_i, N_i, F_i) : -n \le i \le -1)$. Again, we use the bootstrap to compute the corrections $z^*_{A,\ell}, z^*_{B,\ell}, z^*_{A,r}, z^*_{B,r}$ that appear in this setting (that are direct analogs to those appearing in Algorithm 1 for perfect forecasts.)

Algorithm 3

1. Generate Y^*_{-n} as a normal rv with mean $\hat{\mu}/(1-\hat{\rho})$ and variance $\hat{\sigma}^2/(1-\hat{\rho}^2)$.
2. For $-n < i \le -1$, simulate Y_i^* via the recursion

$$Y_i^* = \hat{\rho} Y^*_{i-1} + Z_i^*,$$

where the Z_i^*'s are independently simulated as normal rv's with mean $\hat{\mu}$ and variance $\hat{\sigma}^2$.
3. Given $(Y_i^* : -n \le i \le -1)$, simulate the N_i^*'s as independent Poisson rv's with means $(F_i \exp(Y_i^*) : -n \le i \le -1)$.
4. Compute

$$\hat{M}_2^* = \frac{1}{n}\sum_{i=-n}^{-1}\left(\frac{N_i^{*2} - N_i^*}{F_i^2}\right),$$

$$\hat{M}_3^* = \frac{1}{n-1}\sum_{i=-n+1}^{-1}\left(\frac{N_i^* N_{i-1}^*}{F_i F_{i-1}}\right).$$

5. Compute the minimizer $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\rho}^*)$ of

$$\left(\hat{M}_2^* - m_2(\mu, \sigma^2, \rho)\right)^2 + \left(\hat{M}_3^* - m_3(\mu, \sigma^2, \rho)\right)^2$$

subject to

$$m_1(\mu, \sigma^2, \rho) = 1,$$
$$-1 \le \rho \le 1,$$
$$\sigma^2 \ge 0.$$

6. Generate $(A_i^*, B_i^*, N_i^* - A_i^* - B_i^*)$ as multinomial rv's with parameters $(N_i^*, \hat{p}, \hat{q}, 1-\hat{p}-\hat{q})$, $-n \le i \le -1$.
7. Compute

$$\hat{p}^* = \frac{\sum_{j=-n}^{-1} A_j^*}{\sum_{j=-n}^{-1} N_j^*}$$

$$\hat{q}^* = \frac{\sum_{j=-n}^{-1} B_j^*}{\sum_{j=-n}^{-1} N_j^*}$$

8. Use Algorithm 2 to compute $\ell(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\rho}^*, \hat{p}^* F_r)$, $u(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\rho}^*, \hat{p}^* F_r)$, $\ell(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\rho}^*, \hat{q}^* F_r)$, $u(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\rho}^*, \hat{q}^* F_r)$.
9. Repeat Steps 1 to 8 b times, thereby yielding b 4-tuples $(\ell(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{p}_i^* F_r), u(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{p}_i^* F_r)$, $\ell(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{q}_i^* F_r), u(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{q}_i^* F_r))$.
10. Compute the smallest integers $z^*_{A,\ell}$ and $z^*_{B,\ell}$ for which

$$\frac{1}{b}\sum_{i=1}^{b} I(\ell(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{p}_i^* F_r) - \ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p} F_r) \le z^*_{A,\ell}) \ge 1-\alpha$$

and

$$\frac{1}{b}\sum_{i=1}^{b} I(\ell(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{q}_i F_r) - \ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q} F_r) \le z^*_{B,\ell}) \ge 1-\alpha$$

and the largest integers $z^*_{A,r}, z^*_{B,r}$ for which

$$\frac{1}{b}\sum_{i=1}^{b} I(\ell(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{p}_i^* F_r) - \ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p} F_r) \ge z^*_{A,r}) \ge 1-\alpha$$

and

$$\frac{1}{b}\sum_{i=1}^{b} I(\ell(\hat{\mu}_i^*, \hat{\sigma}_i^{2*}, \hat{\rho}_i^*, \hat{q}_i^* F_r) - \ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q} F_r) \ge z^*_{B,r}) \ge 1-\alpha$$

Then, $[[\ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p} F_r) - z^*_{A,\ell}]^+, u(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p} F_r) - z^*_{A,r}]$ and $[[\ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q} F_r) - z^*_{B,\ell}]^+, u(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q} F_r) - z^*_{B,r}]$ are our bootstrap prediction intervals for A_r and B_r, respectively, based on unbiased log-normal forecasts.

7 Biased forecasts with log-normal errors

We now modify the model of Section 6 to permit biased forecasts. The only change we make here is that we drop

the requirement Eq. 6.2. In this case, we need to add an additional moment identity in order to uniquely identify the coefficients $(\mu_0, \sigma_0^2, \rho_0)$ underlying the forecast errors given by the Γ_i^{-1}'s. Note that

$$\mathbb{E}\left(\frac{N_i}{F_i}\right) = \mathbb{E}N_i(\lambda_i^{-1}\mathbb{E}\Gamma_i) = \mathbb{E}\Gamma_i = m_1(\mu_0, \sigma_0^2, \rho_0).$$

This suggests that we should estimate $(\mu_0, \sigma_0^2, \rho_0)$ via the minimizer $(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})$ of the objective function

$$\sum_{i=1}^{3}\left(\hat{M}_i - m_i(\mu, \sigma^2, \rho)\right)^2 \tag{7.1}$$

subject to

$$-1 \le \rho \le 1,$$
$$\sigma^2 \ge 0,$$

where $\hat{M}_2$ and $\hat{M}_3$ are defined as in Section 4 and

$$\hat{M}_1 = \sum_{i=-n}^{-1}\left(\frac{N_i}{F_t}\right).$$

As in Section 4, we estimate p_0 and q_0 via $\hat{p}$ and $\hat{q}$ as in Eq. 4.2. As in Section 4, $[\ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p}F_r), u(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p}F_r)]$ and $[\ell(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q}F_r), u(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{q}F_r)]$ are then our plug-in prediction intervals for N_r based on the biased log-normal forecast error model.

Similarly, incorporating the estimation error related to estimating $(\mu_0, \sigma_0^2, \rho_0, p_0, q_0)$ via $(\hat{\mu}, \hat{\sigma}^2, \hat{\rho}, \hat{p}, \hat{q})$ requires only small modifications to the methodology of Section 6. The modified version of Algorithm 3 reflecting use of biased forecasts is provided next.

Algorithm 4 Algorithm 4 is identical to Algorithm 3, excepting that $(\hat{\mu}, \hat{\sigma}^2, \hat{\rho})$ is now the minimizer of Eq. 7.1, and Steps 4 and 5 are modified as follows:

4′. Compute

$$\hat{M}_1^* = \frac{1}{n}\sum_{i=-n}^{-1}\left(\frac{N_i^*}{F_i}\right),$$
$$\hat{M}_2^* = \frac{1}{n}\sum_{i=-n}^{-1}\left(\frac{N_i^{*2} - N_i^*}{F_i^2}\right),$$
$$\hat{M}_3^* = \frac{1}{n-1}\sum_{i=-n+1}^{-1}\left(\frac{N_i^* N_{i-1}^*}{F_i F_{i-1}}\right).$$

5′. Compute the minimizer $(\hat{\mu}^*, \hat{\sigma}^{2*}, \hat{\rho}^*)$ of

$$\sum_{i=1}^{3}\left(\hat{M}_i - m_i(\mu, \sigma^2, \rho)\right)^2$$

subject to

$$-1 \le \rho \le 1,$$
$$\sigma^2 \ge 0.$$

Algorithm 4 yields our desired bootstrap prediction intervals for A_r and B_r, just as Algorithm 3 yields such intervals for the unbiased model of Section 6.

8 Use at a large academic medical center and evaluation using synthetic data

8.1 Model deployment and evaluation: empirical data

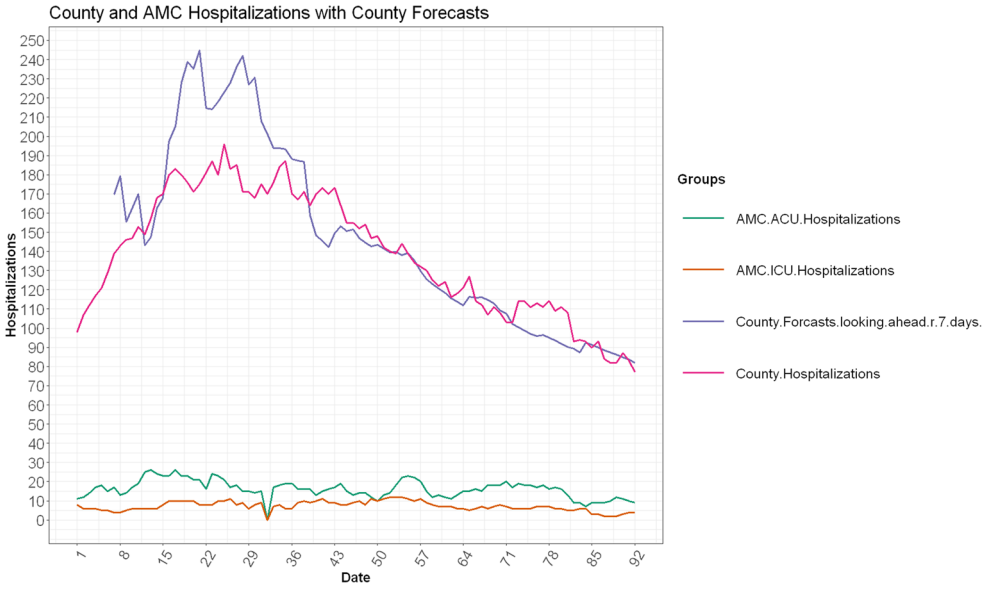
We use historical county-level COVID-19 hospitalization forecasts and ACU, ICU COVID-19 hospitalizations from the AMC studied. Given the small number of patients, we protect patient privacy by replacing the actual date with the number of days from a reference date during the summer of 2020. The values are shown in Fig. 1.

We compare the prediction intervals under the three settings we discuss above (perfect, unbiased, biased), with plug-in prediction intervals and bootstrap prediction intervals under each setting. We choose $\delta = 0.05$ (corresponding to 95% prediction intervals) for both plug-in and bootstrap, and $\alpha = 0.05$ (corresponding to confidence level of 95%) for bootstrap.

To compare the prediction performance of each proposed model with real data, we choose $r = 7$ (on each Monday we make ACU, ICU predictions for the next Monday), comparing with the actual value. Also we set n increased by 1 for each additional observed day in each model. We set algorithm parameters $b_0 = 1000, m = 300, \delta = 0.05, \alpha = 0.05$. With these parameters, the perfect model, unbiased model and biased model proposed converged in 1.25, 4.08 and 4.25 seconds, respectively, when they were run by R 3.6.3 on a computer with an Intel Core i7-1065G7 (4 cores, 8 processors) and 32 GB RAM.

Projections for ACU and ICU made by different models are shown in Figs. 2 and 3. In each plot, dash lines indicate 95% bootstrap prediction intervals, solid lines indicate 95% plug-in prediction intervals and black dots indicate actual values.

Fig. 1 Empirical Data



The unbiased models tend to provide wider prediction intervals. As we get larger n, the bootstrap prediction intervals are getting closer to the plugin prediction intervals.

As shown in Table 1, with $r = 7$, the fractions of weeks for which each 95% plug-in prediction intervals covered the observed bed count in the ACU is 70% for all three models. The 95% bootstrap prediction intervals covered 90% of the observed bed count in the ACU for all models. All of the prediction intervals covered 100% of the observed bed count in the ICU. The results with 90% prediction intervals (Figs. 12, 13 and Table 4) and 80% prediction intervals (Figs. 14, 15 and Table 5) on AMC data can be found in the Appendix.

The demand intervals forecast using the perfect model were communicated to the hospital manager in charge of COVID-19 response capacity planning. The upper bound

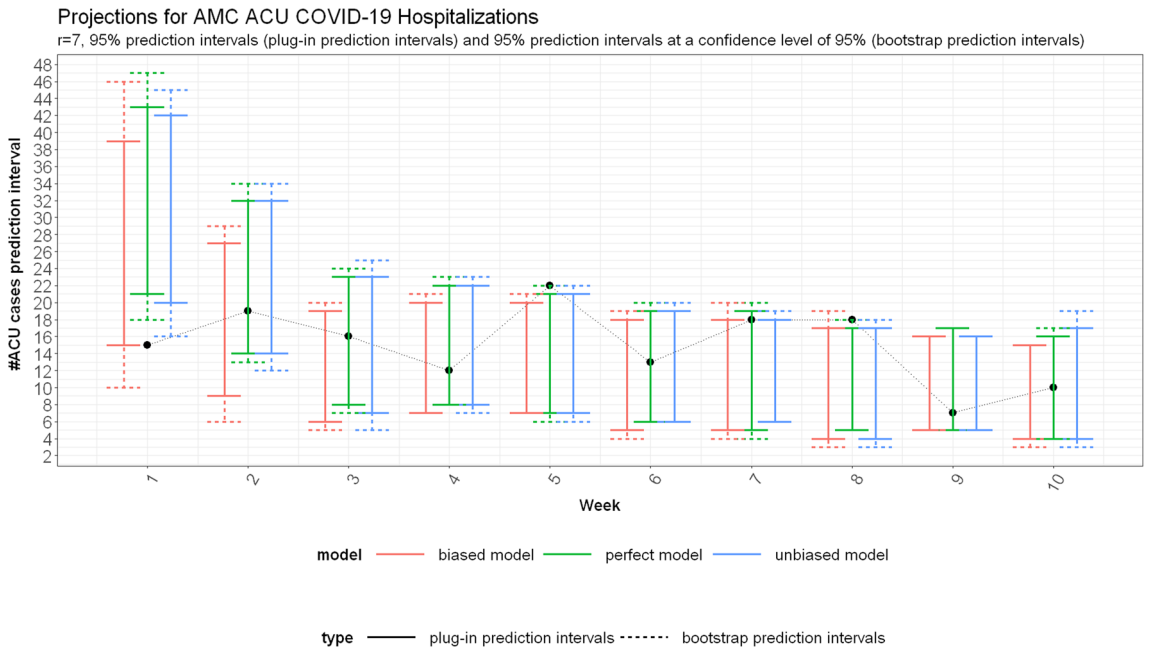


Fig. 2 AMC ACU projections, $r = 7$, 95% prediction intervals, with black dots representing actual values

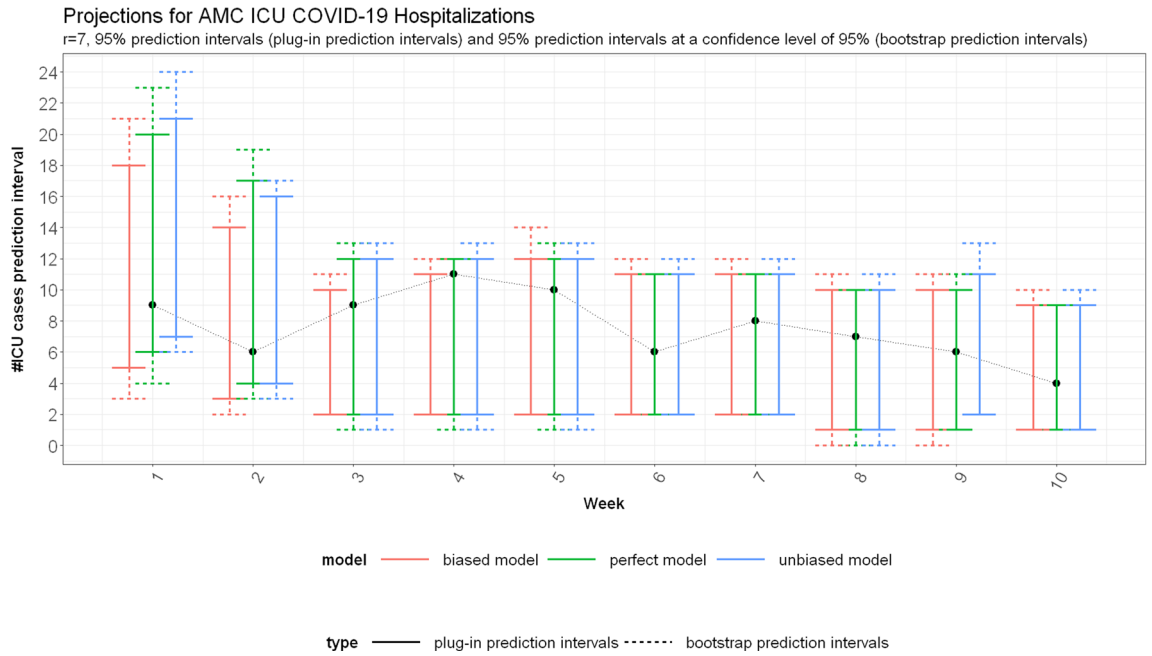


Fig. 3 AMC ICU projections, $r = 7$, 95% prediction intervals, with black dots representing actual values

of the prediction intervals remained below the threshold hospital leadership felt comfortable could be accommodated without the cancellation of elective admissions.

8.2 Performance evaluation: synthetic data

In this section, we generate two sets of synthetic data for 100 days. In both examples, we generate N_i from the Poisson distribution with mean λ_i. A_i's and B_i's are generated from the multinomial distribution with parameters $(N_i, p, q, 1-p-q)$. They are all generated once and used in all following sections.

In Example 1, the λ's are generated using SIR model (see, for example, [39] and [40]). In particular, here we set the initial number of infections as 5, the initial population as 1000, the infection rate $\alpha = 0.2$ and the recovery rate $\gamma = 0.1$. In this example, we set the ratio of patients hospitalized in ACU and ICU at hospital level as $p = 0.14, q = 0.05$ respectively.

In Example 2, we generate λ's uniformly on the supports changing by time. For day 1 to day 20, the λ's are uniformly generated from {100, 101,..., 149, 150}; for day 21 to day 50, the λ's are uniformly generated from {20, 21, ..., 99, 100}; in the last 50 days, the λ's are uniformly generated from {100, 101, ..., 199, 200}. In this example, we set the ratio of patients hospitalized in ACU and ICU at hospital level as $p = 0.5, q = 0.2$ respectively.

The synthetic forecasts are generated following different model assumptions. To evaluate the performances, we apply the above prediction methods on the last 60 observations. The synthetic data are shown in Figs. 4 and 5.

Table 1 Coverage rate of 95% plug-in prediction intervals and 95% bootstrap prediction intervals at a confidence level of 95%, AMC

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	70%	90%	100%	100%
Unbiased Model	70%	90%	100%	100%
Biased Model	70%	90%	100%	100%

8.2.1 Synthetic data under perfect forecasts model

Here we generate

$$F_i = \lambda_i, i = 1, ..., 100$$

satisfying the "perfect forecast" assumption. The 95% prediction intervals for ACU($\{A_i\}$), ICU($\{B_i\}$) are shown in Figs. 6 (Example 1) and 7 (Example 2).

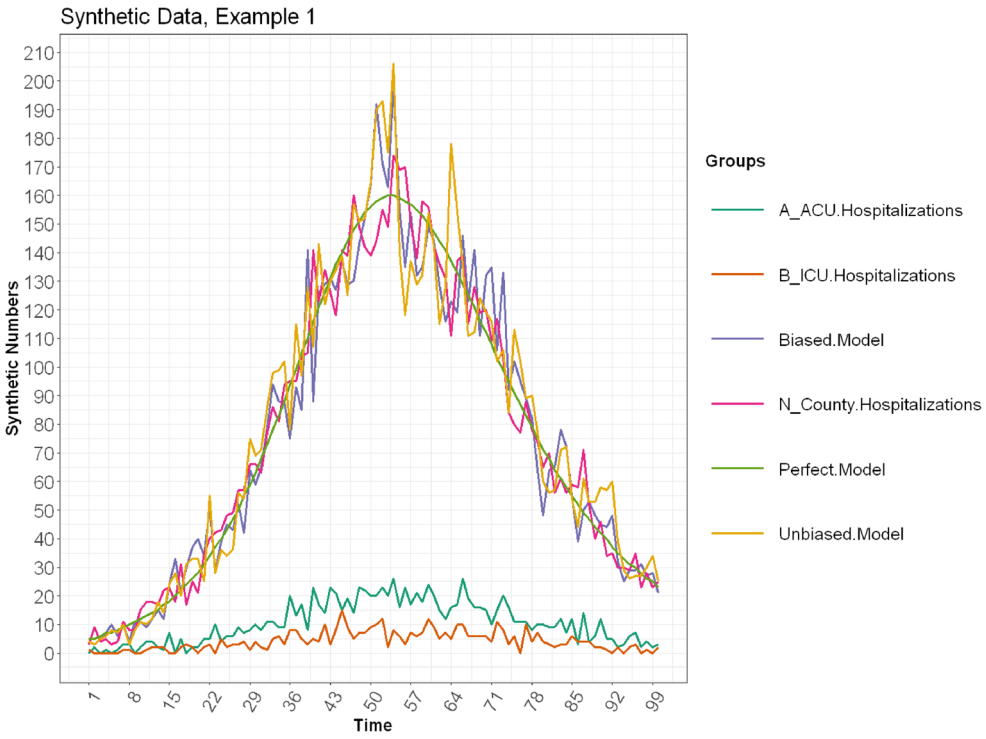


Fig. 4 Synthetic Data, Example 1

The fractions of observations for which 95% plug-in prediction intervals covered the observed bed count are 97%, 92% for ACU, ICU respectively in Example 1, and the ones for Example 2 are 92% and 92%; the ones for 95% bootstrap prediction intervals are 98%, 98% in Example 1 and 98%, 97% in Example 2.

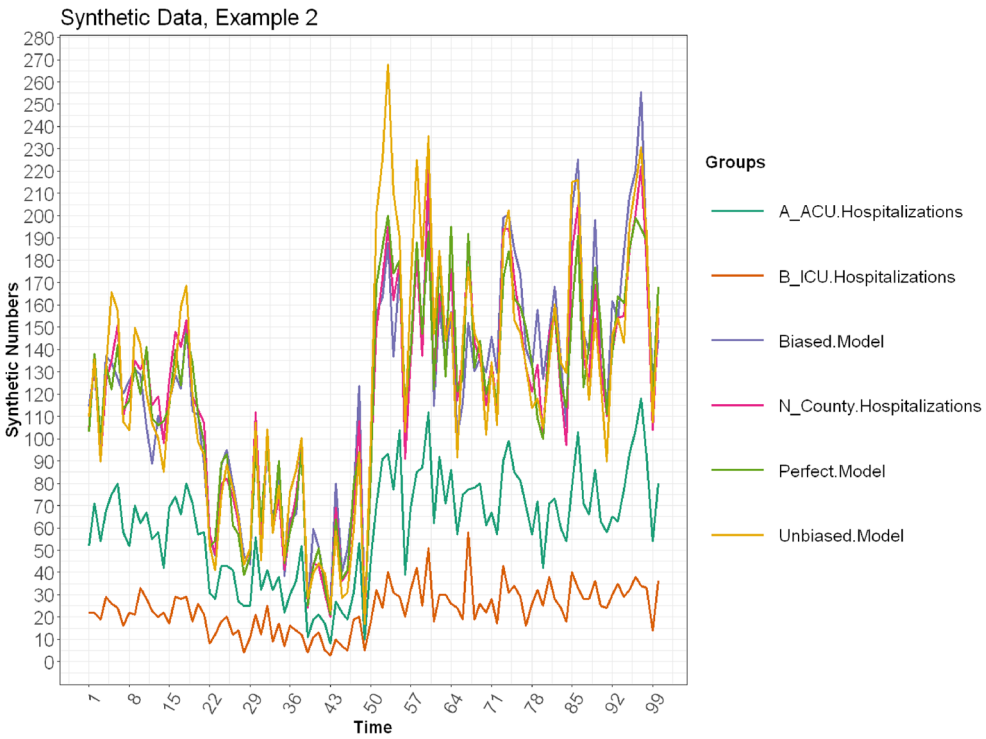


Fig. 5 Synthetic Data, Example 2

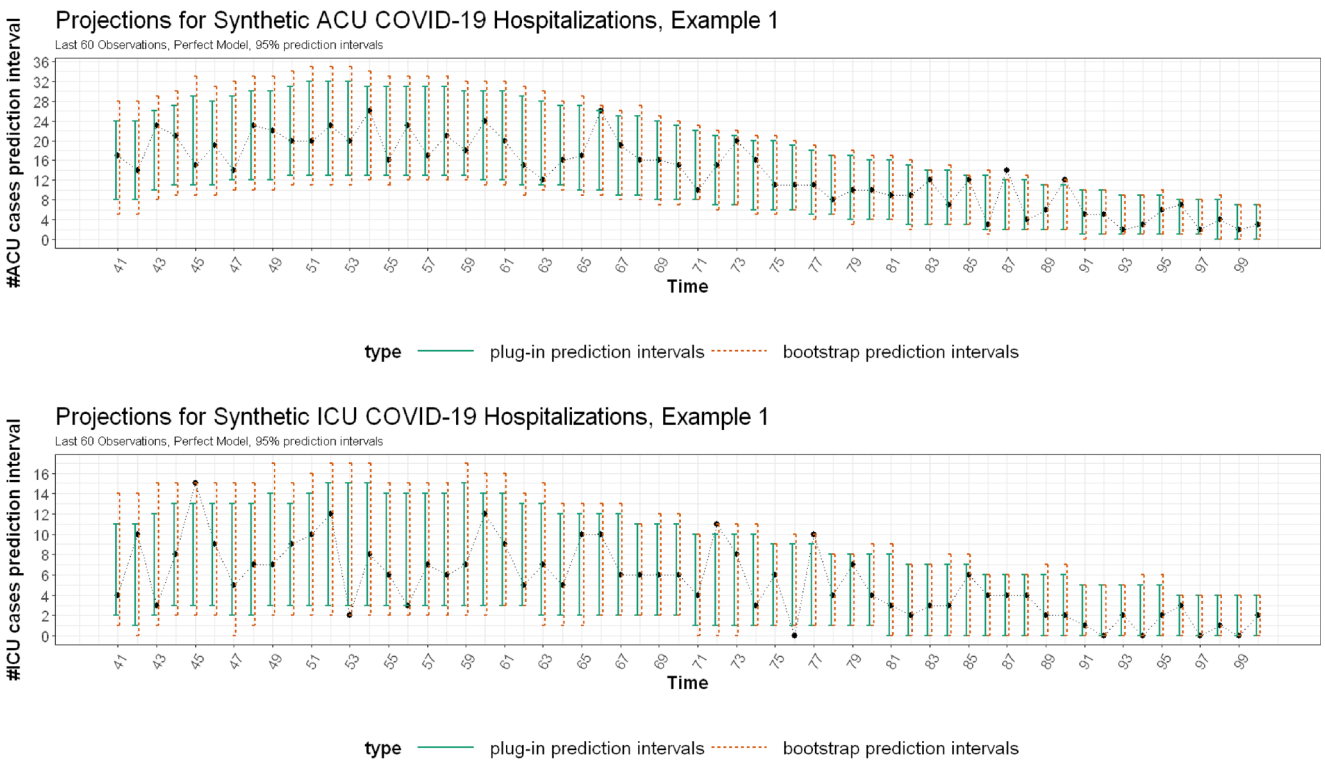


Fig. 6 Projections on synthetic data, Example 1, 95% prediction intervals, perfect model, with black dots representing actual values

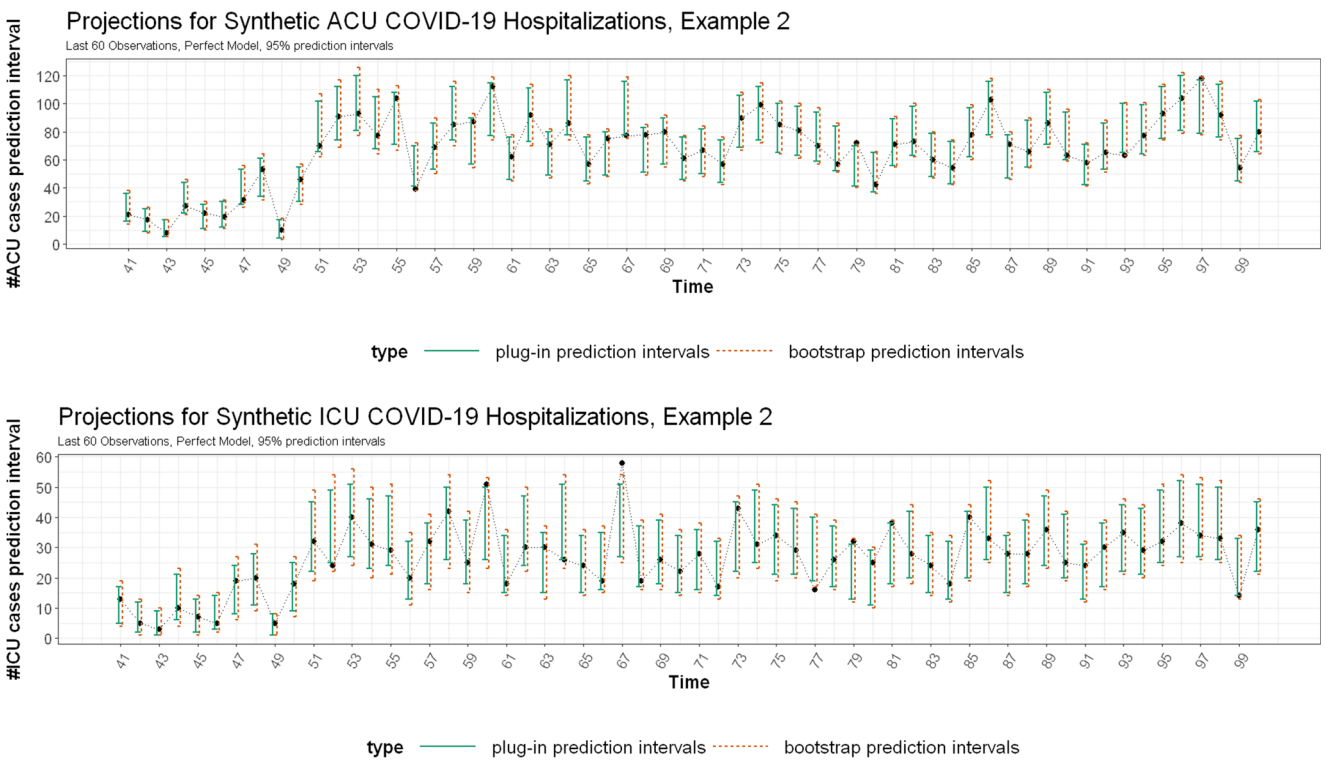


Fig. 7 Projections on synthetic data, Example 2, 95% prediction intervals, perfect model, with black dots representing actual values

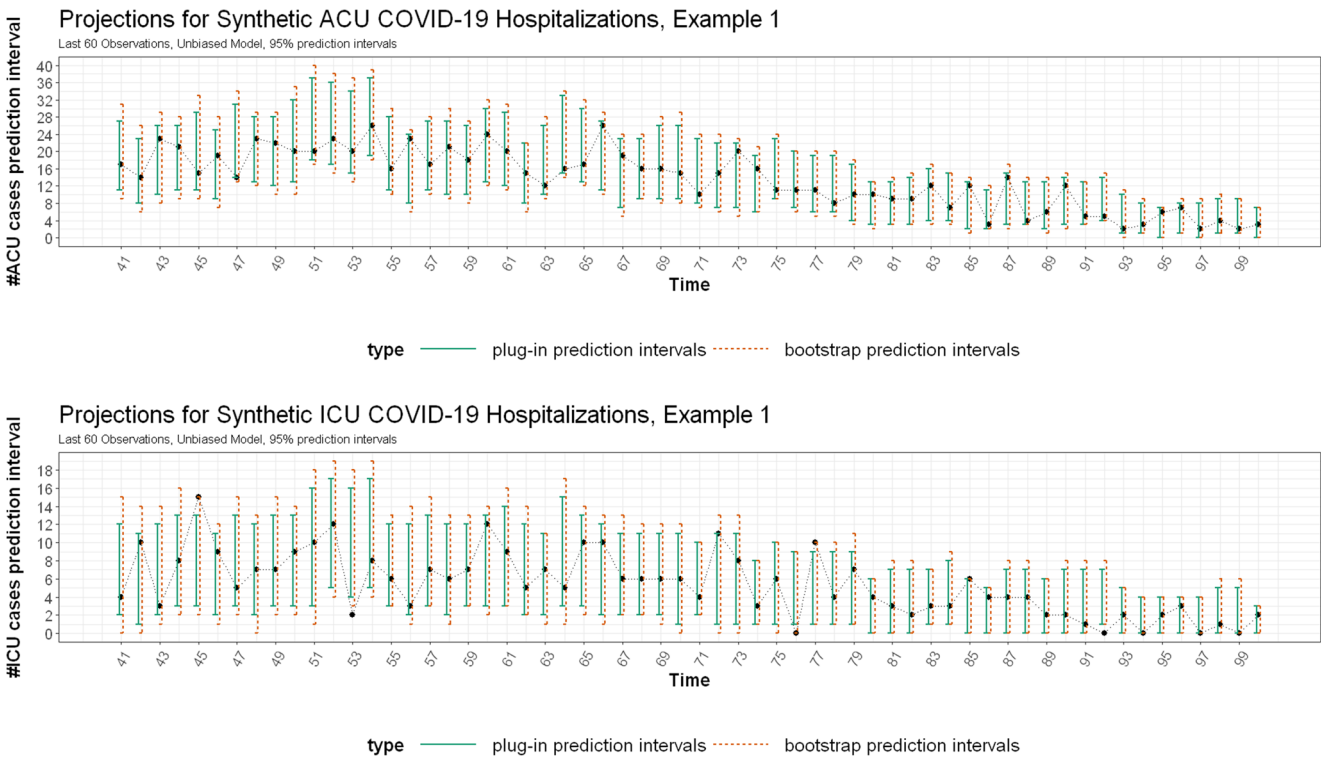


Fig. 8 Projections on synthetic data, Example 1, 95% prediction intervals, unbiased model, with black dots representing actual values

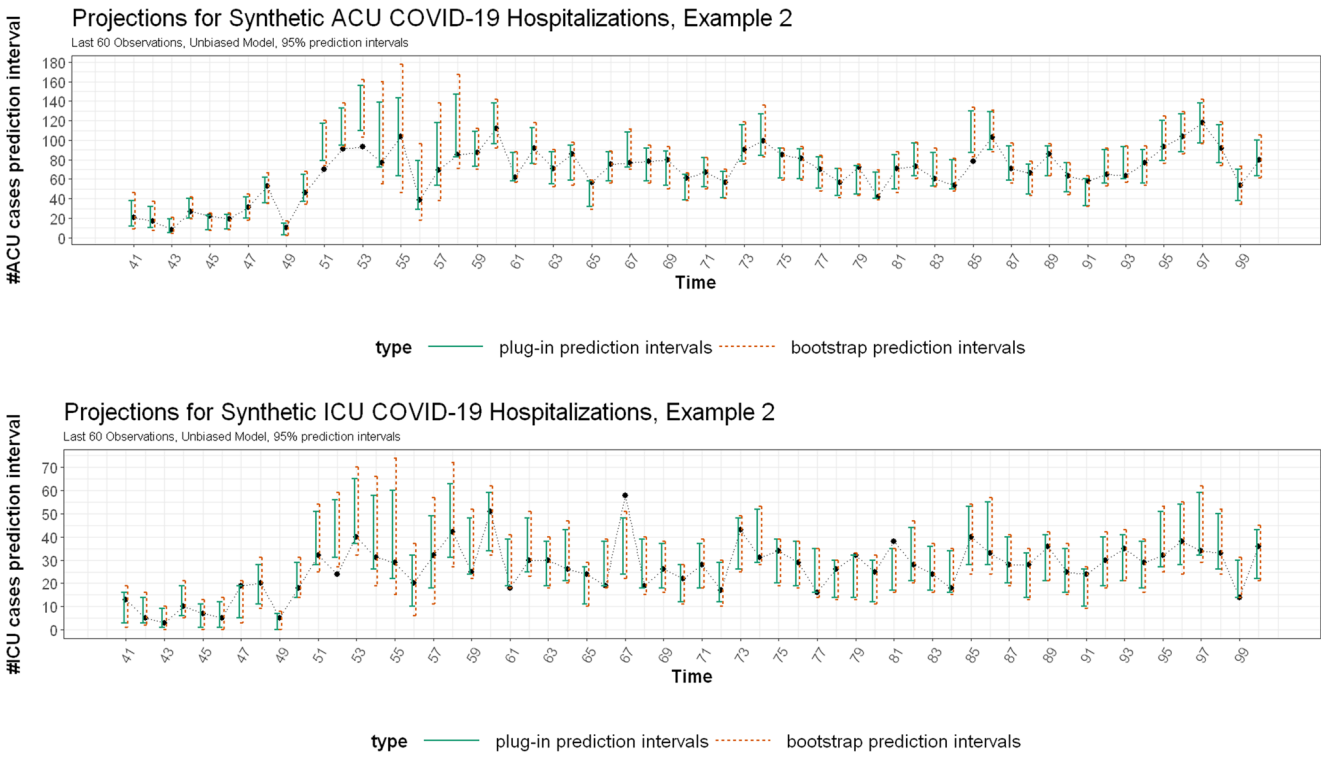


Fig. 9 Projections on synthetic data, Example 2, 95% prediction intervals, unbiased model, with black dots representing actual values

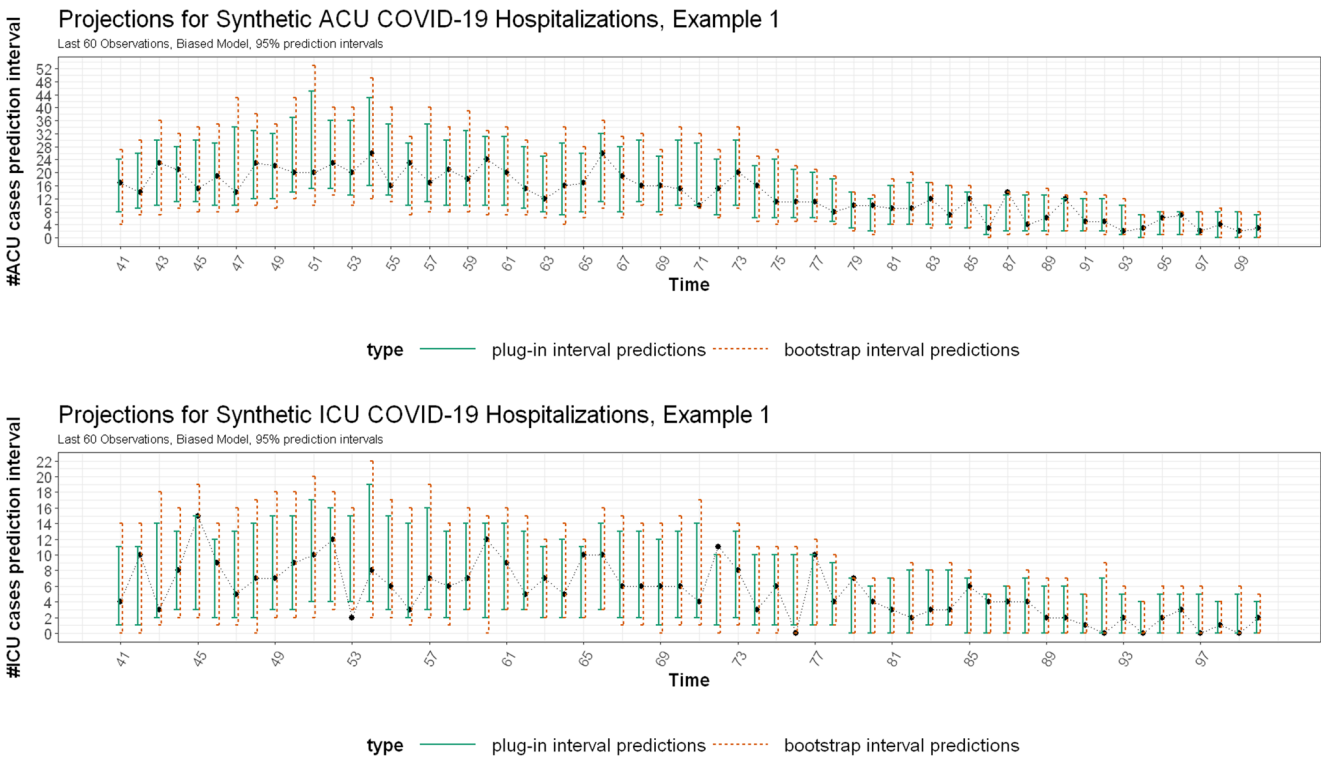


Fig. 10 Projections on synthetic data, Example 1, 95% prediction intervals, biased model, with black dots representing actual values

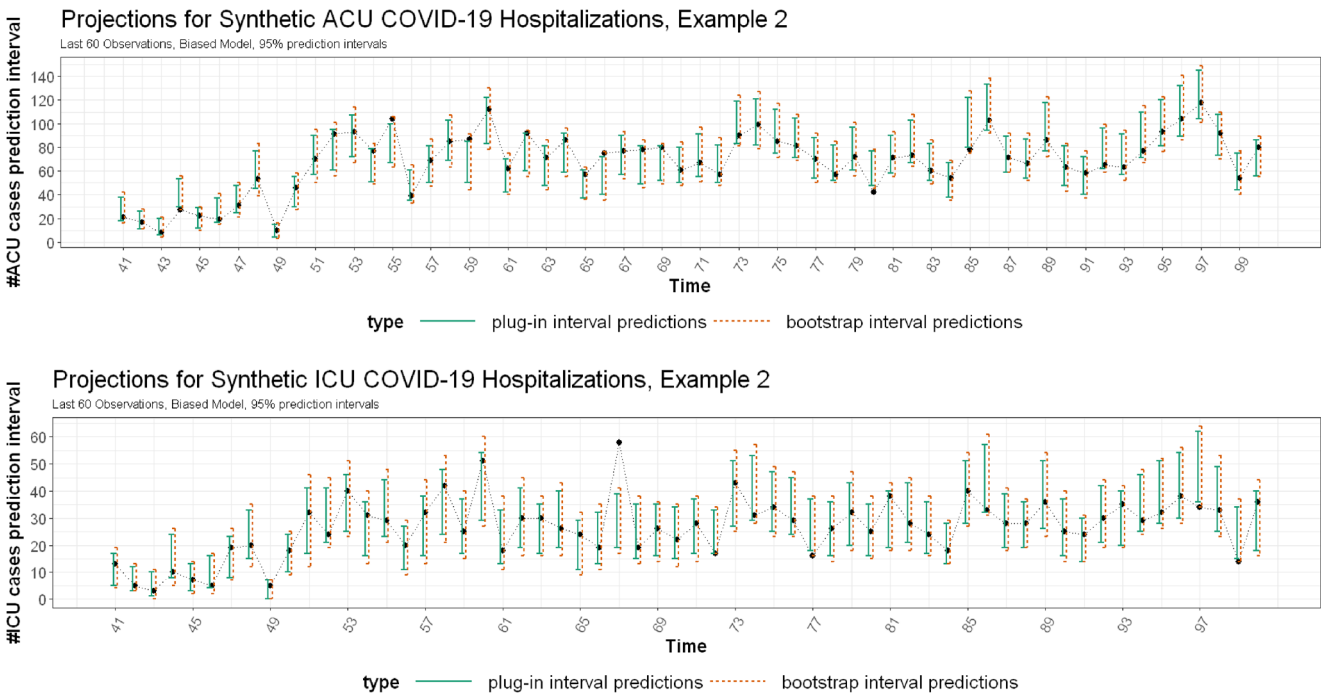


Fig. 11 Projections on synthetic data, Example 2, 95% prediction intervals, biased model, with black dots representing actual values

Table 2 Coverage rate of 95% prediction intervals, synthetic data, Example 1

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	97%	98%	92%	98%
Unbiased Model	100%	100%	92%	97%
Biased Model	98%	100%	95%	97%

8.2.2 Synthetic data under unbiased forecasts model

Under this setting, we set $\rho = 0.5, \sigma^2 = 0.01, \mu = -\frac{\sigma^2}{2(1+\rho)}$ and generate

$$Y_1 \sim N\left(\frac{\mu}{1-\rho}, \frac{\sigma^2}{1-\rho^2}\right)$$

$$Y_i = \rho Y_{i-1} + Z_i, Z_i \sim N(\mu, \sigma^2), i = 2, ..., 100$$

$$\Gamma_i = exp(Y_i), F_i = \frac{\lambda_i}{\Gamma_i}, i = 1, ..., 100,$$

where $\sim$ represents "distributed according to", so that $\mathbb{E}(\Gamma_i) = 1$ which satisfies the assumptions in the "unbiased forecasts model". The 95% prediction intervals for ACU($\{A_i\}$), ICU($\{B_i\}$) of the two examples are shown in Figs. 8 and 9. The fractions of observations for which 95% plug-in prediction intervals covered the observed bed count are 100%, 92% for ACU, ICU respectively in Example 1, and the fractions for Example 2 are 93% and 93%; the ones for 95% bootstrap prediction intervals are 100% and 97% in Example 1, and 93%, 95% in Example 2.

8.2.3 Synthetic data under biased forecasts model

Under this setting, we set $\rho = 0.5, \sigma^2 = 0.01, \mu = 0$. The generation method is the same as that in unbiased model setting. The 95% prediction intervals for ACU($\{A_i\}$), ICU($\{B_i\}$) are shown in Figs. 10 and 11. The fractions of observations for which plug-in prediction intervals covered the observed bed count are 98%, 95% for ACU, ICU respectively in Example 1, and the fractions for Example 2 are 90% and 93%; the ones for bootstrap prediction intervals are 100%, 97% in Example 1, and 98%, 98% in Example 2.

Table 3 Coverage rate of 95% prediction intervals, synthetic data, Example 2

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	92%	98%	92%	97%
Unbiased Model	93%	93%	93%	95%
Biased Model	90%	98%	93%	98%

All the plots show that with n increasing, the bootstrap prediction intervals are getting closer to the plugin intervals. The coverage rates of 95% prediction intervals on the two examples are shown in Tables 2 and 3. The results with 90% prediction intervals (Figs. 16, 17, 18 and Table 6 for Example 1, Figs. 19, 20, 21 and Table 7 for Example 2) and 80% prediction intervals (Figs. 22, 23, 24 and Table 8 for Example 1, Figs. 25, 26, 27 and Table 9 for Example 2) on both sets of synthetic data can be found in the Appendix.

9 Conclusions

In this work we introduce, DICE (Demand Intervals from Consistent Estimators), a model to forecast prediction intervals for the fraction of regional patient demand arriving to an institution based on the historical fraction of demand served by the institution and, potentially biased, forecasts of demand as a Poisson random variable. We show that our model is consistent, computationally tractable, and well-calibrated on real-world data as well as synthetic data. Unlike other flu-specific or general forecasting models in the literature, our model produces integral prediction, principled intervals around these estimates even for small values of the estimate and does so using only three data sources. The use of regional-level forecasts, that are commonly available and incorporate numerous population-specific considerations, allows the model to take advantage of rich contextual data without increasing the complexity of its implementation or reducing its generalizability. The calibration of the model is evidenced by evaluation on real-world and synthetic data as the intervals generated narrow as uncertainty is removed from the inputs and cover the observations approximately the percentage of the time that they are expected to. To illustrate its potential usefulness, we discuss the managerial COVID-19 decisions that prompted the development of the models as well as how they were used to inform these decisions at an academic medical center. The demand interval forecasts suggested that the "second wave" influx of COVID-19 patients would be unlikely to exceed available hospital capacity. The information provided by the model contributed to, the ultimately correct, decision that COVID-19 patients could be accommodated without the cancellation of elective admissions.

Over the course of the pandemic, numerous hospitals went from seeing relatively few patients to being overwhelmed with new arrivals relatively quickly. Even relatively large confidence intervals, such as when relatively little historical data are available, may reassure hospital decision makers compared to the alternative of potentially unbounded exponential growth in arrivals. The extended evaluation of the model with synthetic data shows that it is well calibrated, i.e. when sufficient data are available the prediction intervals are appropriately sized.

This work has several limitations and opportunities for subsequent research. The present model does not account for scenarios in which the total demand for hospital beds approaches the available capacity of the region. Subsequent work is necessary to expand the model to capture the fixed total capacity of hospitals in a region and the routing of patients from hospitals at capacity to hospitals with capacity available. The present model does not examine how demand forecasts and uncertainty intervals are translated into operational decision making. Subsequent work should examine how, for example, the forecast can be used to estimate patient load in the next few days, if the staff that is scheduled is sufficient, and options for cancelling or rescheduling procedures more dynamically than at a fixed cadence of 14 days. Another area for further research is the use of DICE for patient demand unrelated to COVID-19. Other areas of urgent and non-urgent surgical and medical demand that change as the standard of care or composition of the population changes may be subject to this type of forecasting if relatively reliable regional forecasts are available.

As hospitals the world over prepare for a third wave of COVID-19, this model may find similar applications at institutions planning their response to an influx of patients. Beyond COVID-19, patient demand for a variety of medical conditions is forecast as a Poisson random variable. DICE may be of use to the numerous decision to make which hospital managers project demand for their services by combining their historical share of regional demand with forecasts of total regional demand.

Appendix

We establish here that the "method of moments" estimators of Sections 4 and 5 will be consistent in great generality. This will follow if we can prove that

$$\hat{M}_i \xrightarrow{p} m_i(\mu_0, \sigma_0^2, \rho_0) \tag{1}$$

as $n \to \infty$, for $1 \le i \le 3$, where $\xrightarrow{p}$ denotes "converge in probability".

Note that $\mathbb{E}\hat{M}_i = m_i(\mu_0, \sigma_0^2, \rho_0)$ for $1 \le i \le 3$. Hence, Eq. 1 follows from Chebyshev's inequality if we can show that

$$\mathrm{Var}\hat{M}_i \to 0$$

as $n \to \infty$. But

$$\begin{aligned}\mathrm{Var}\hat{M}_1 &= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=-n}^{-1}\frac{N_i}{F_i}\right)\\ &= \frac{1}{n^2}\sum_{i=-n}^{-1}\mathrm{Var}\left(\frac{N_i}{F_i}\right)\\ &\quad+\frac{2}{n^2}\sum_{i=-n}^{-1}\sum_{j=i+1}^{-n}\mathrm{Cov}\left(\frac{N_i}{F_i},\frac{N_j}{F_j}\right)\end{aligned}$$

Of course,

$$\begin{aligned}\mathrm{Var}\left(\frac{N_i}{F_i}\right) &= \mathbb{E}\left(\frac{N_i^2}{F_i^2}\right)-\left(\mathbb{E}\left(\frac{N_i}{F_i}\right)\right)^2\\ &= \frac{\mathbb{E}N_i^2}{\lambda_i^2}\cdot\mathbb{E}\Gamma_i^2 - m_1(\mu_0,\sigma_0^2,\rho_0)^2\\ &= \frac{\lambda_i+\lambda_i^2}{\lambda_i^2}\cdot\mathbb{E}\Gamma_{-1}^2\\ &= \mathrm{Var}\Gamma_{-1}+\frac{1}{\lambda_i}\mathbb{E}\Gamma_{-1}^2.\end{aligned}$$

Also, for $i < j$

$$\begin{aligned}\mathrm{Cov}\left(\frac{N_i}{F_i},\frac{N_j}{F_j}\right) &= \mathbb{E}\left(\frac{N_i}{F_i}\frac{N_j}{F_j}\right)-\mathbb{E}\left(\frac{N_i}{F_i}\right)\mathbb{E}\left(\frac{N_j}{F_j}\right)\\ &= \frac{\mathbb{E}N_i\mathbb{E}N_j}{\lambda_i\lambda_j}\mathbb{E}\Gamma_i\Gamma_j - m_1(\mu_0,\sigma_0^2,\rho_0)^2\\ &= \mathrm{Cov}\left(\Gamma_{-1-(j-i)},\Gamma_{-1}\right)\\ &= \mathbb{E}\exp(Y_{-1-(j-i)}+Y_{-1})-(\mathbb{E}\Gamma_{-1})^2\\ &= \mathbb{E}\exp((1+\rho^{-j-i})Y_{-1-(j-i)}\\ &\qquad+\sum_{j=0}^{j-i-1}\rho^j Z_{-1-j})-(\mathbb{E}\Gamma_{-1})^2\\ &= O(\rho^{j-i}),\end{aligned}$$

Table 4 Coverage rate of 90% prediction intervals, AMC

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	60%	60%	100%	100%
Unbiased Model	60%	60%	100%	100%
Biased Model	70%	90%	100%	100%

Table 5 Coverage rate of 80% prediction intervals, AMC

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	60%	60%	90%	90%
Unbiased Model	60%	70%	100%	100%
Biased Model	60%	70%	90%	90%

where $O(a_i)$ denotes a quantity that is bounded by a multiple of $|a_i|$. Similar calculations can be found in, for example, [41]. Consequently,

$$\begin{aligned}\mathrm{Var}\hat{M}_1 &= \frac{1}{n^2}\sum_{i=-n}^{-1}\left[\mathrm{Var}\Gamma_{-1}+\frac{1}{\lambda_i}\mathbb{E}\Gamma_{-1}^2\right]+\\ &\quad\frac{2}{n^2}\sum_{i=-n}^{-1}\sum_{j=i+1}^{-n}O(\rho^{j-i})\\ &= O\left(\frac{1}{n}\right)\to 0\end{aligned}$$

as $n \to \infty$ if we assume that the infimum of the λ_i's is bounded away from zero. Similarly, $\mathrm{Var}\hat{M}_i \to 0$ for $i = 2$ and $i = 3$ under this very moderate hypothesis on the λ_i's, thereby establishing the consistency.

Table 6 Coverage rate of 90% prediction intervals, synthetic data, Example 1

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	95%	95%	90%	92%
Unbiased Model	97%	98%	90%	92%
Biased Model	95%	98%	92%	97%

Table 7 Coverage rate of 90% prediction intervals, synthetic data, Example 2

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	92%	92%	87%	92%
Unbiased Model	90%	93%	93%	93%
Biased Model	87%	90%	88%	90%

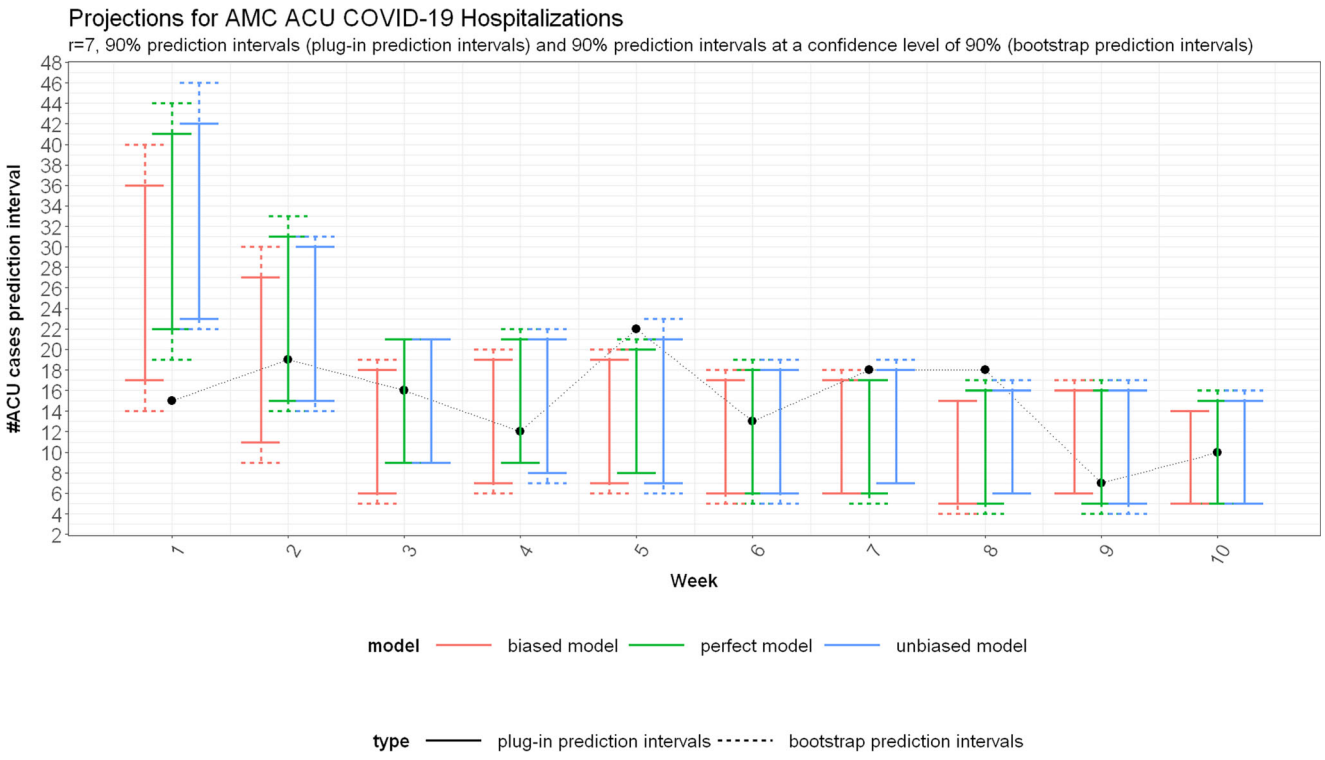


Fig. 12 AMC ACU projections, $r = 7$, 90% prediction intervals, with black dots representing actual values

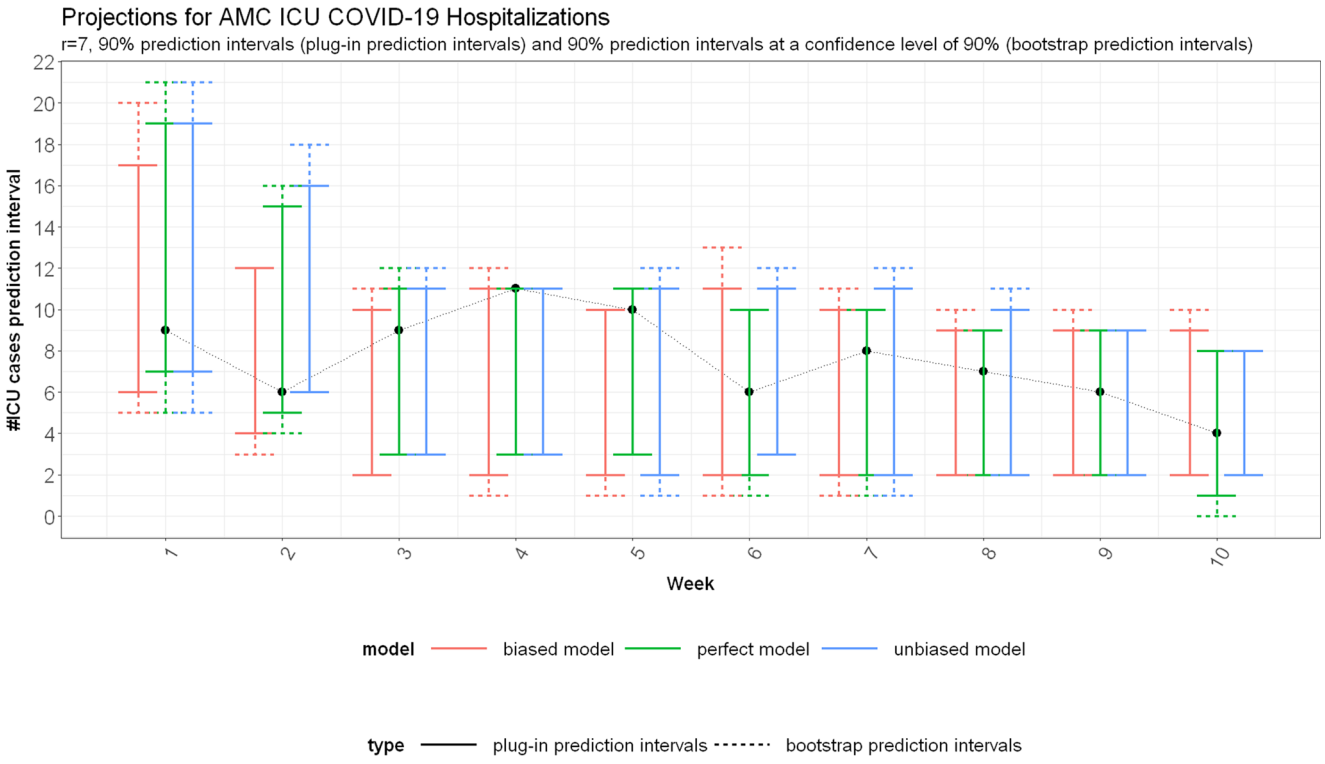


Fig. 13 AMC ICU projections, $r = 7$, 90% prediction intervals, with black dots representing actual values

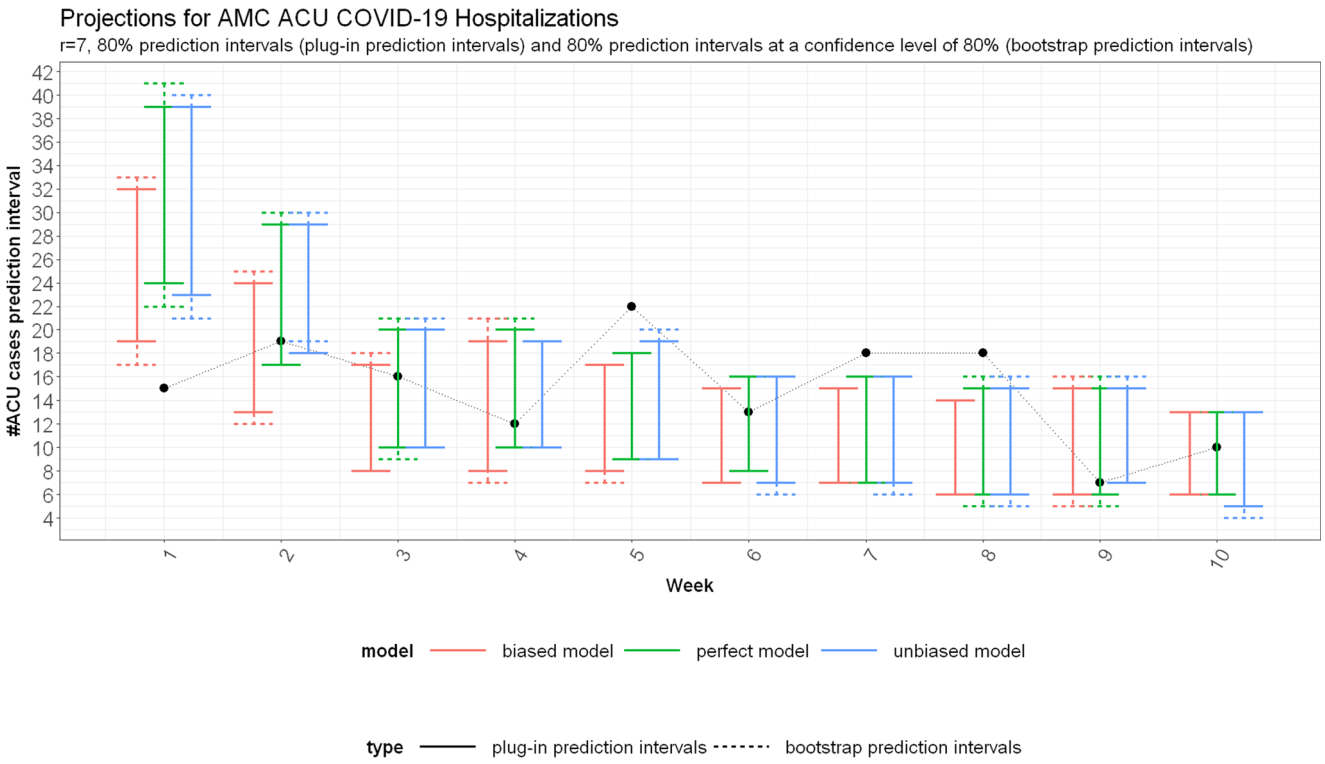


Fig. 14 AMC ACU projections, $r = 7$, 80% prediction intervals, with black dots representing actual values

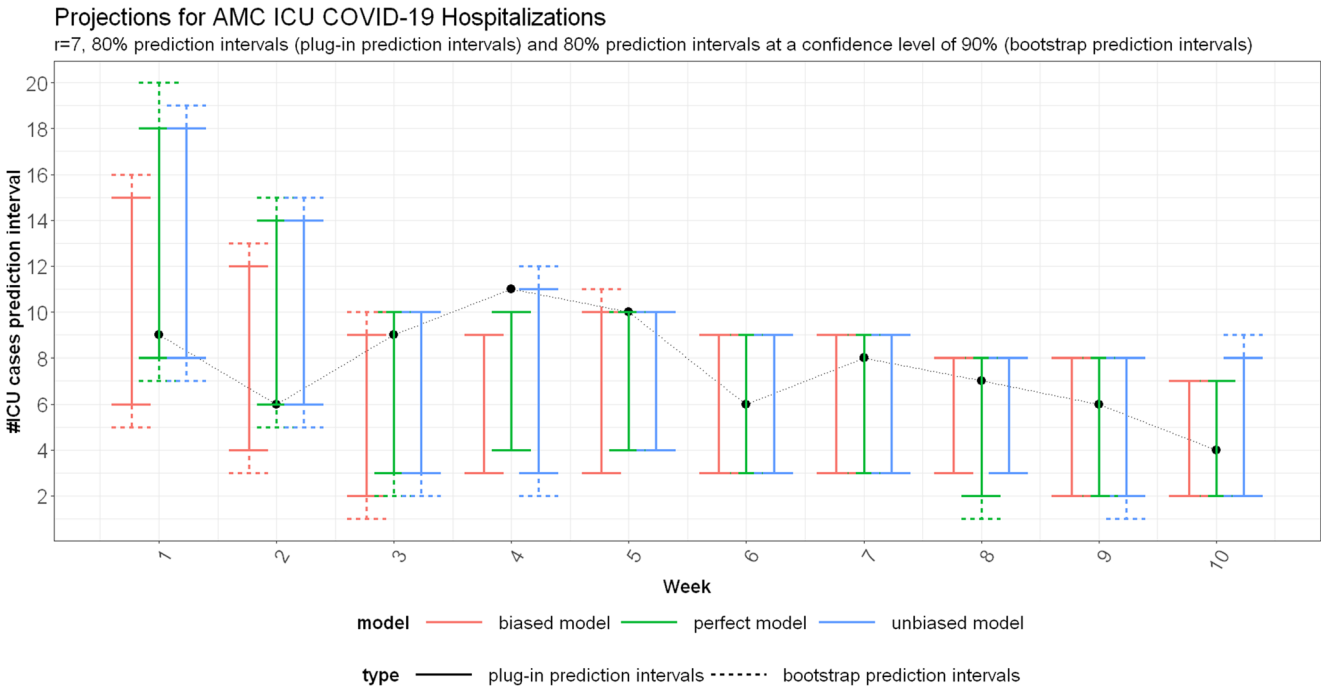


Fig. 15 AMC ICU projections, $r = 7$, 80% prediction intervals, with black dots representing actual values

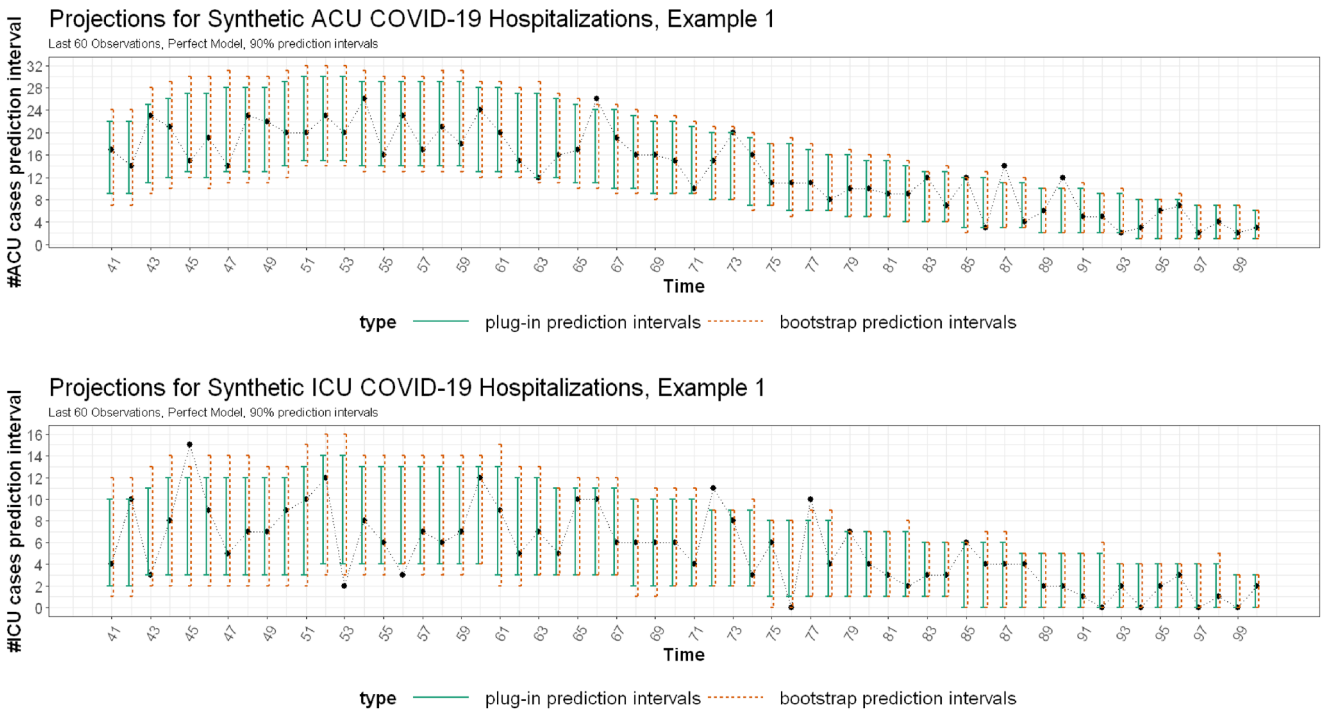


Fig. 16 Projections on synthetic data, Example 1, 90% prediction intervals, perfect model, with black dots representing actual values

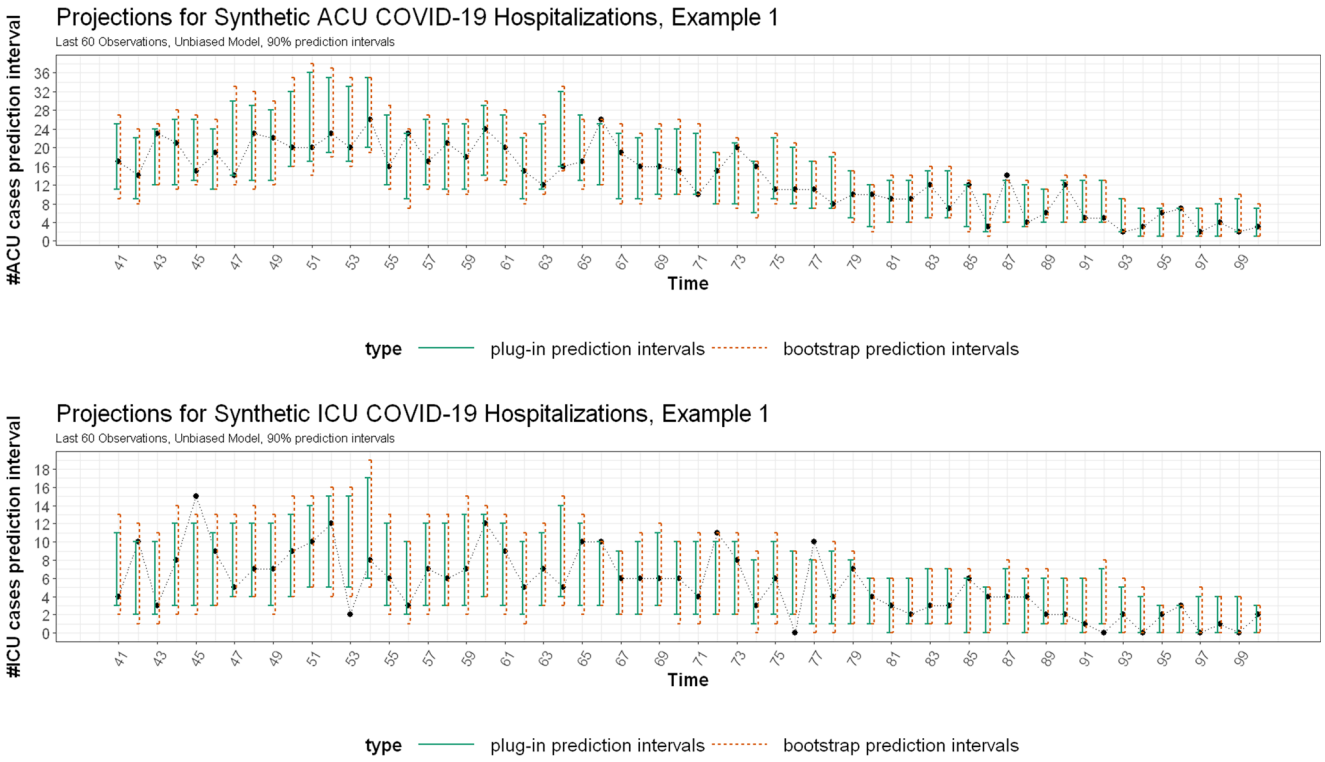


Fig. 17 Projections on synthetic data, Example 1, 90% prediction intervals, biased model, with black dots representing actual values

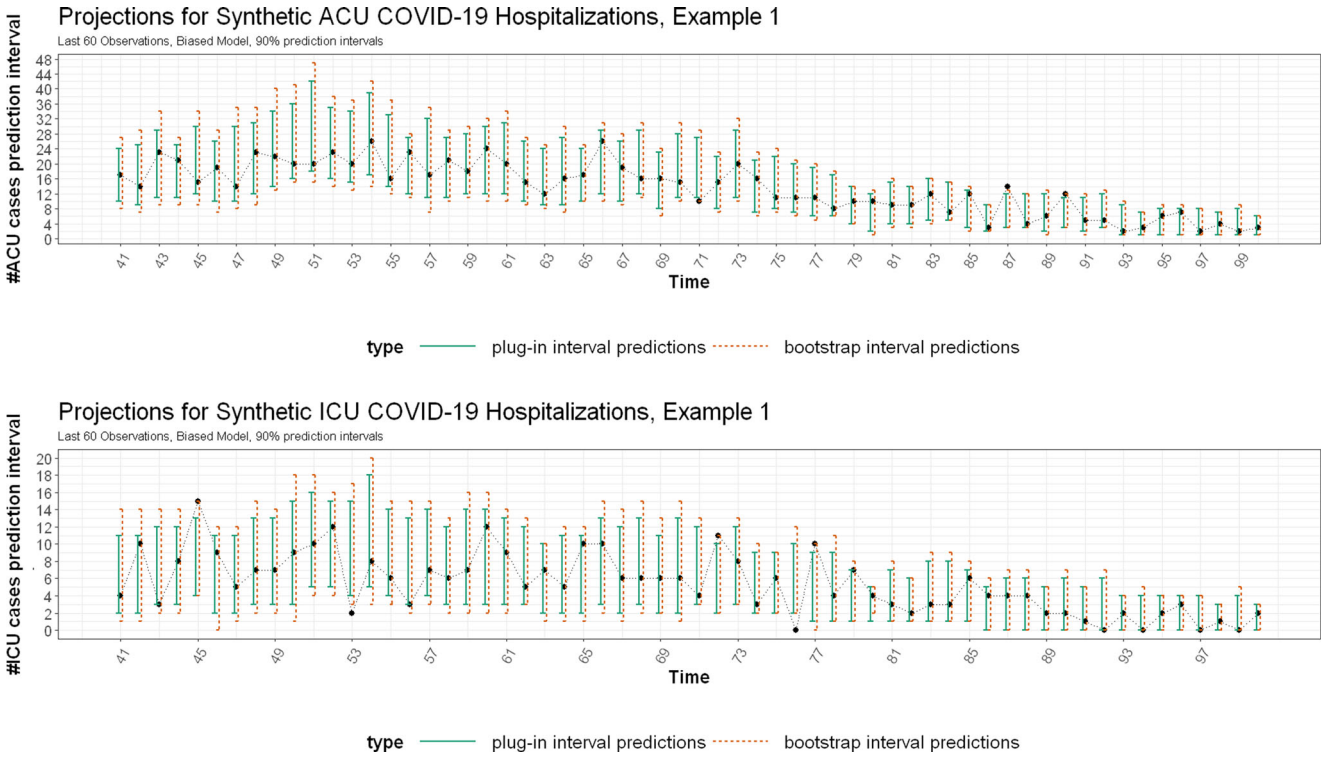


Fig. 18 Projections on synthetic data, Example 1, 90% prediction intervals, biased model, with black dots representing actual values

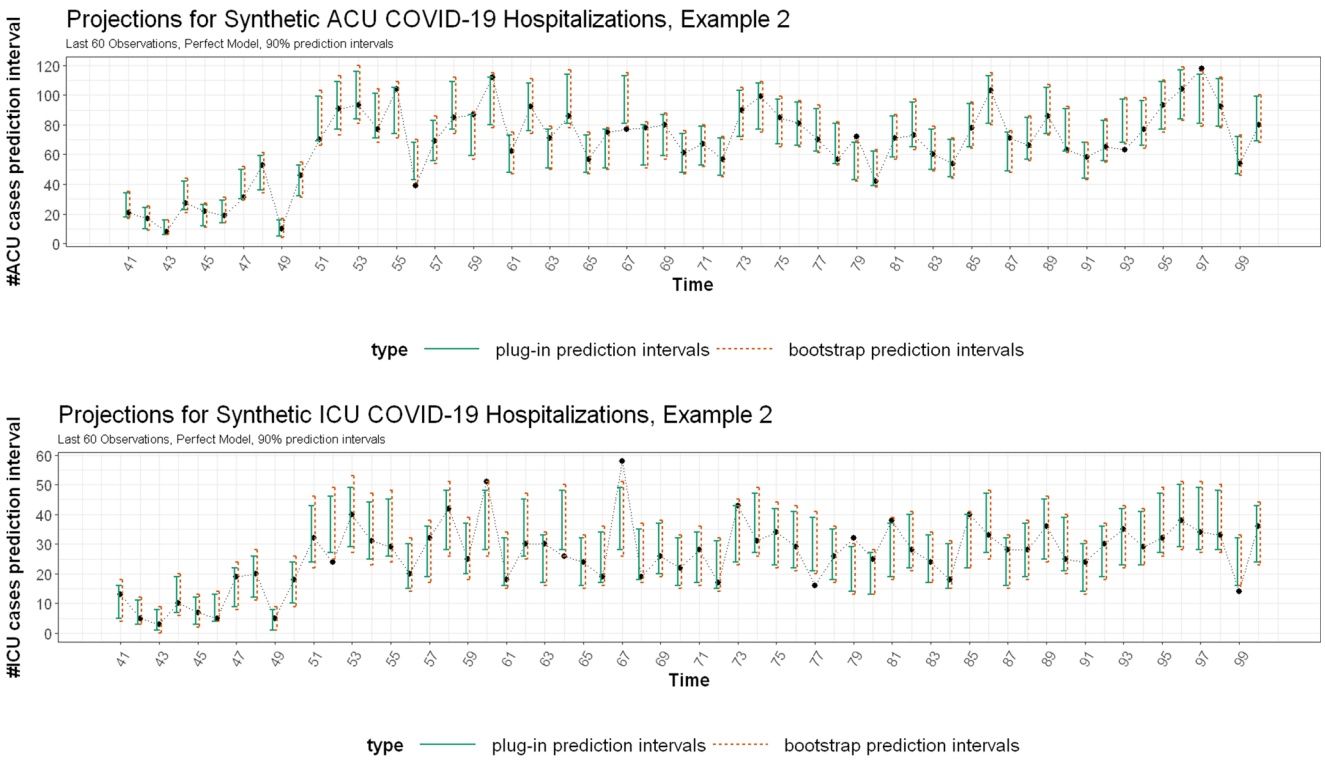


Fig. 19 Projections on synthetic data, Example 2, 90% prediction intervals, perfect model, with black dots representing actual values

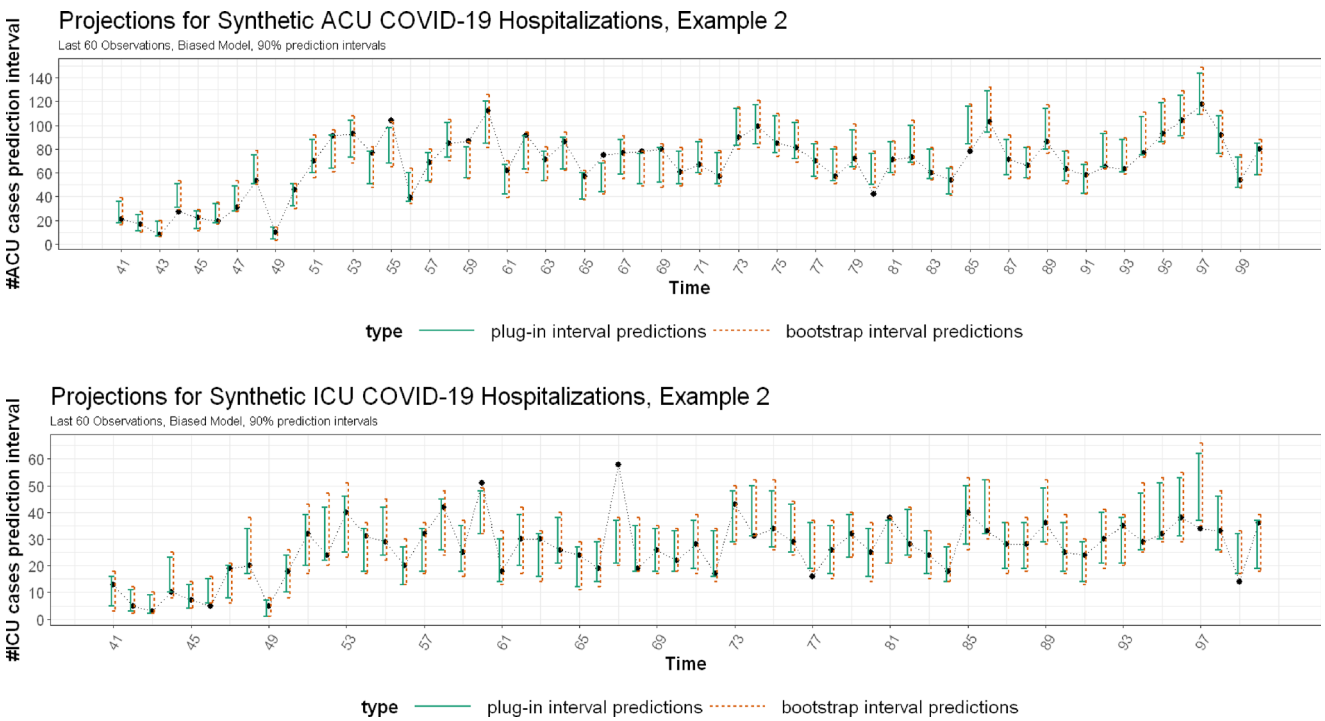


Fig. 20 Projections on synthetic data, Example 2, 90% prediction intervals, biased model, with black dots representing actual values

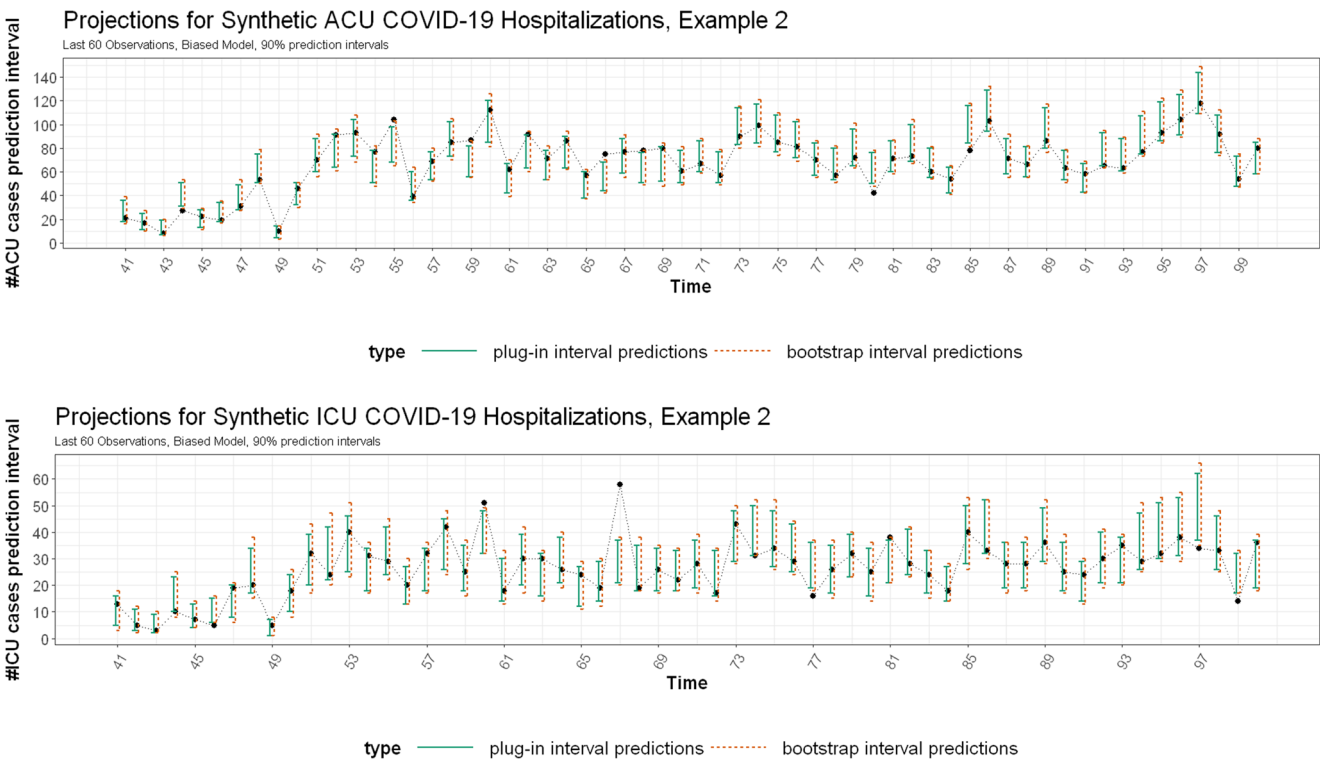


Fig. 21 Projections on synthetic data, Example 2, 90% prediction intervals, biased model, with black dots representing actual values

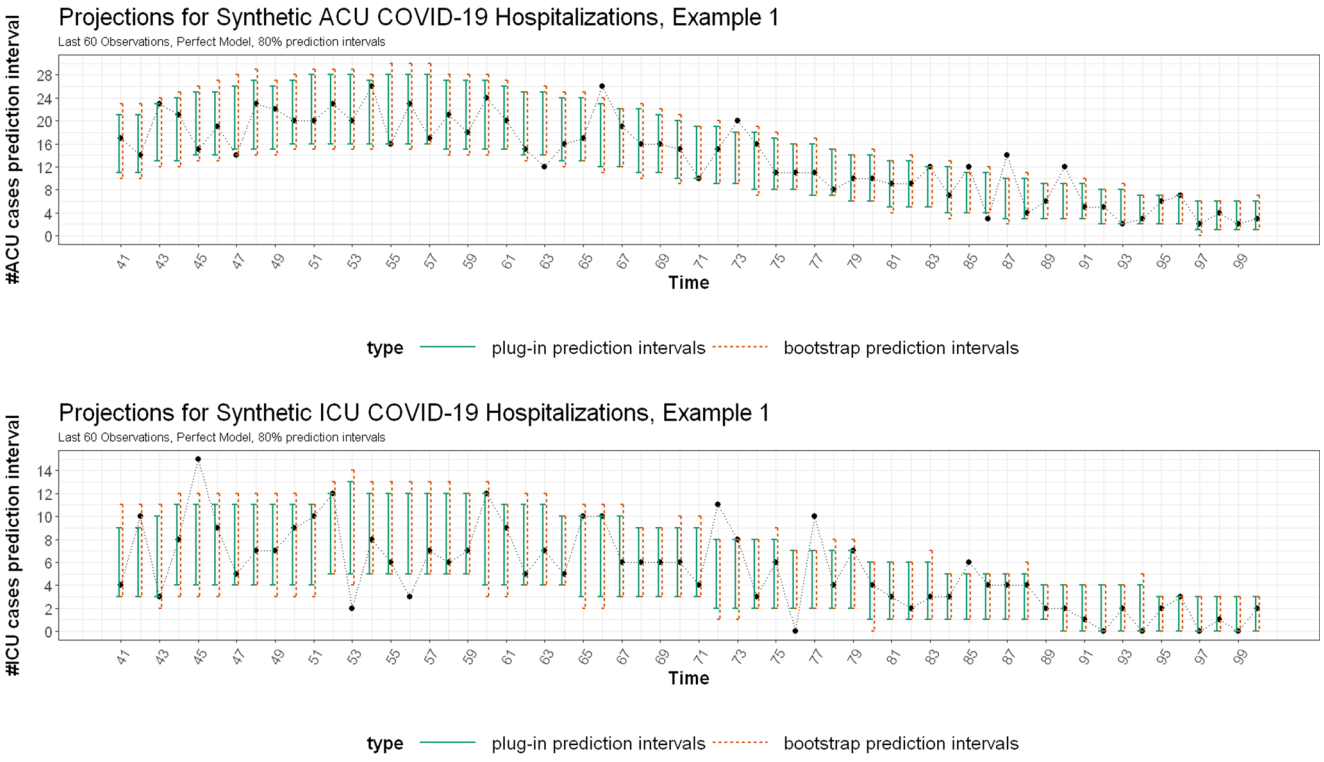


Fig. 22 Projections on synthetic data, Example 1, 80% prediction intervals, perfect model, with black dots representing actual values

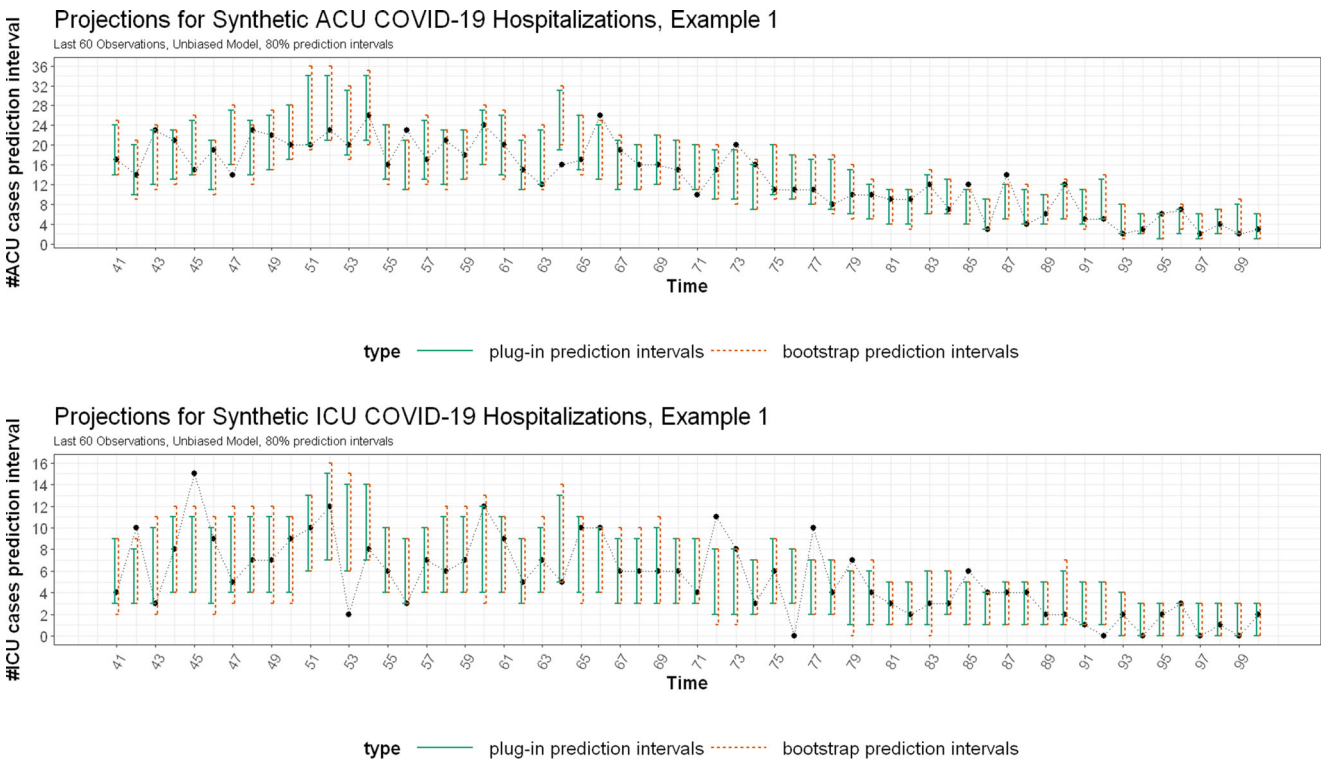


Fig. 23 Projections on synthetic data, Example 1, 80% prediction intervals, unbiased model, with black dots representing actual values

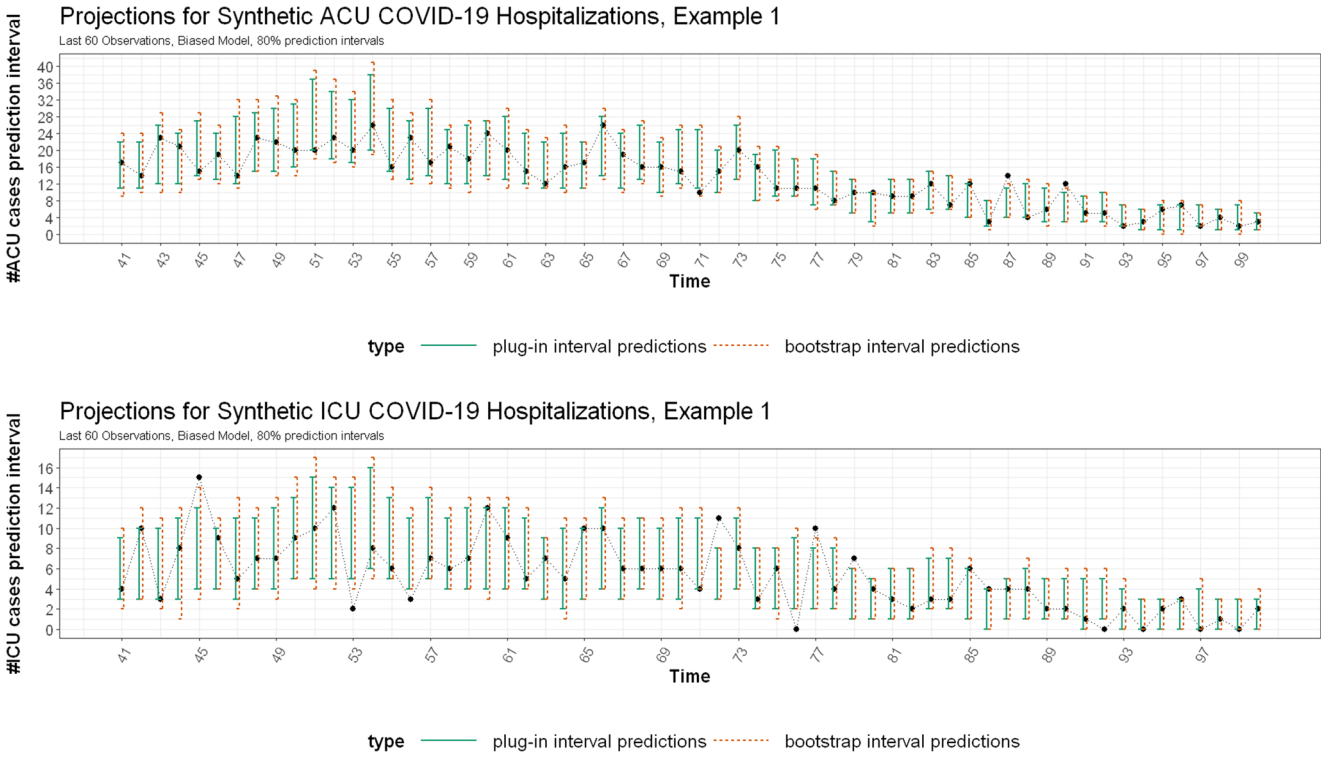


Fig. 24 Projections on synthetic data, Example 1, 80% prediction intervals, biased model, with black dots representing actual values

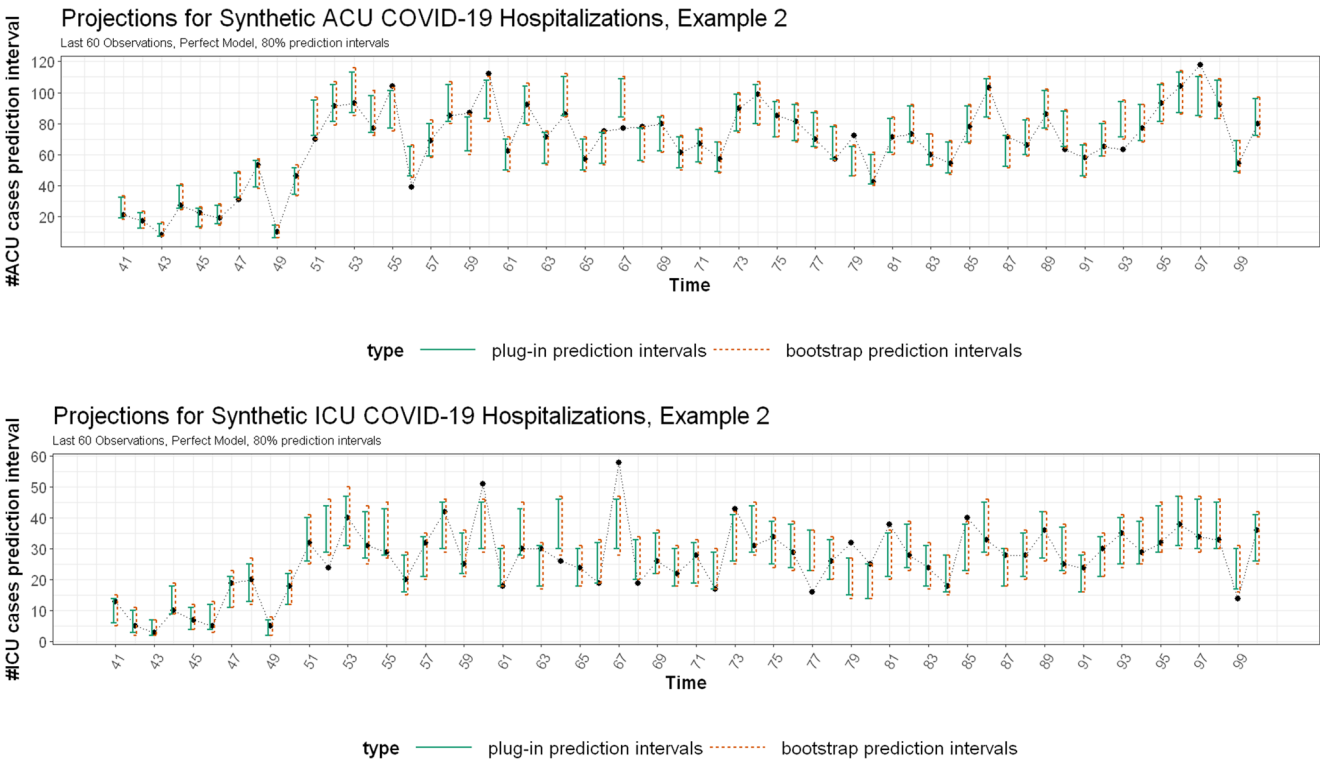


Fig. 25 Projections on synthetic data, Example 2, 80% prediction intervals, perfect model, with black dots representing actual values

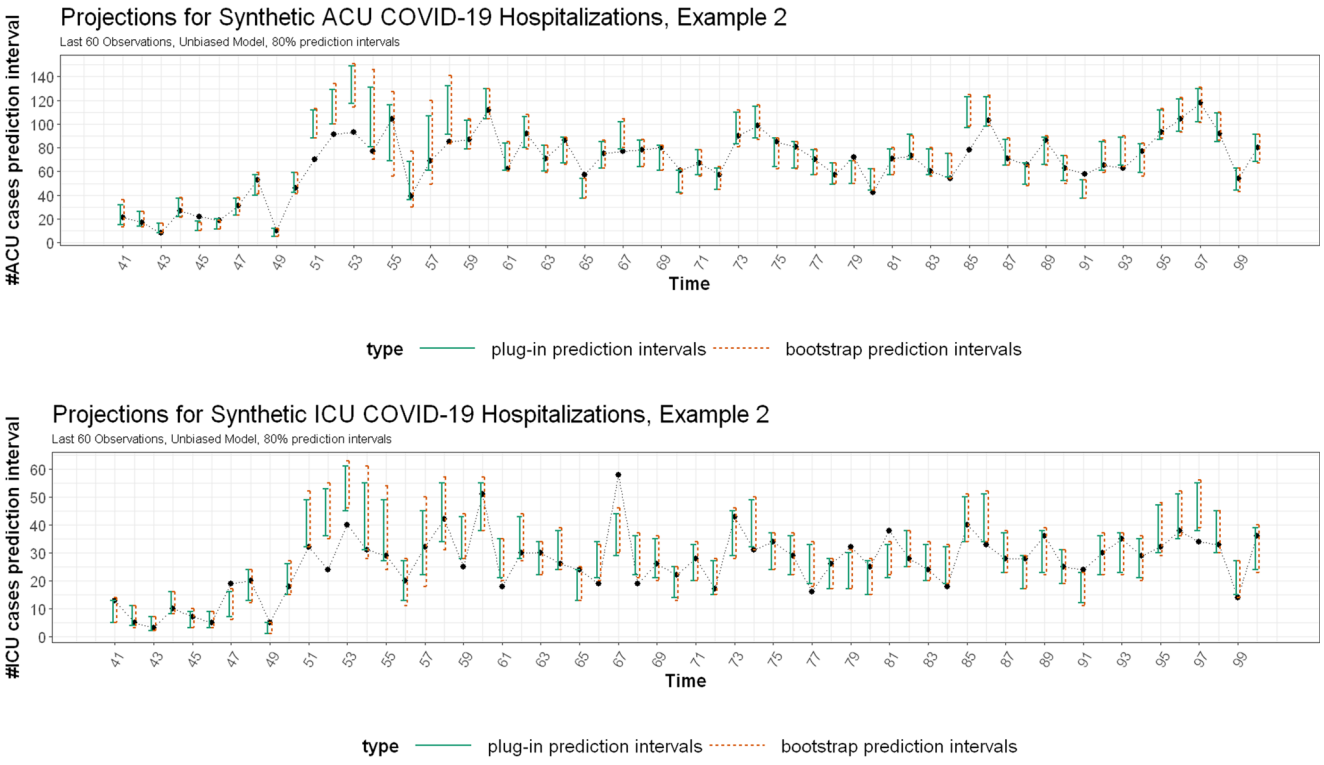


Fig. 26 Projections on synthetic data, Example 2, 80% prediction intervals, unbiased model, with black dots representing actual values

Table 8 Coverage rate of 80% prediction intervals, synthetic data, Example 1

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	87%	88%	87%	88%
Unbiased Model	87%	87%	85%	85%
Biased Model	95%	97%	87%	87%

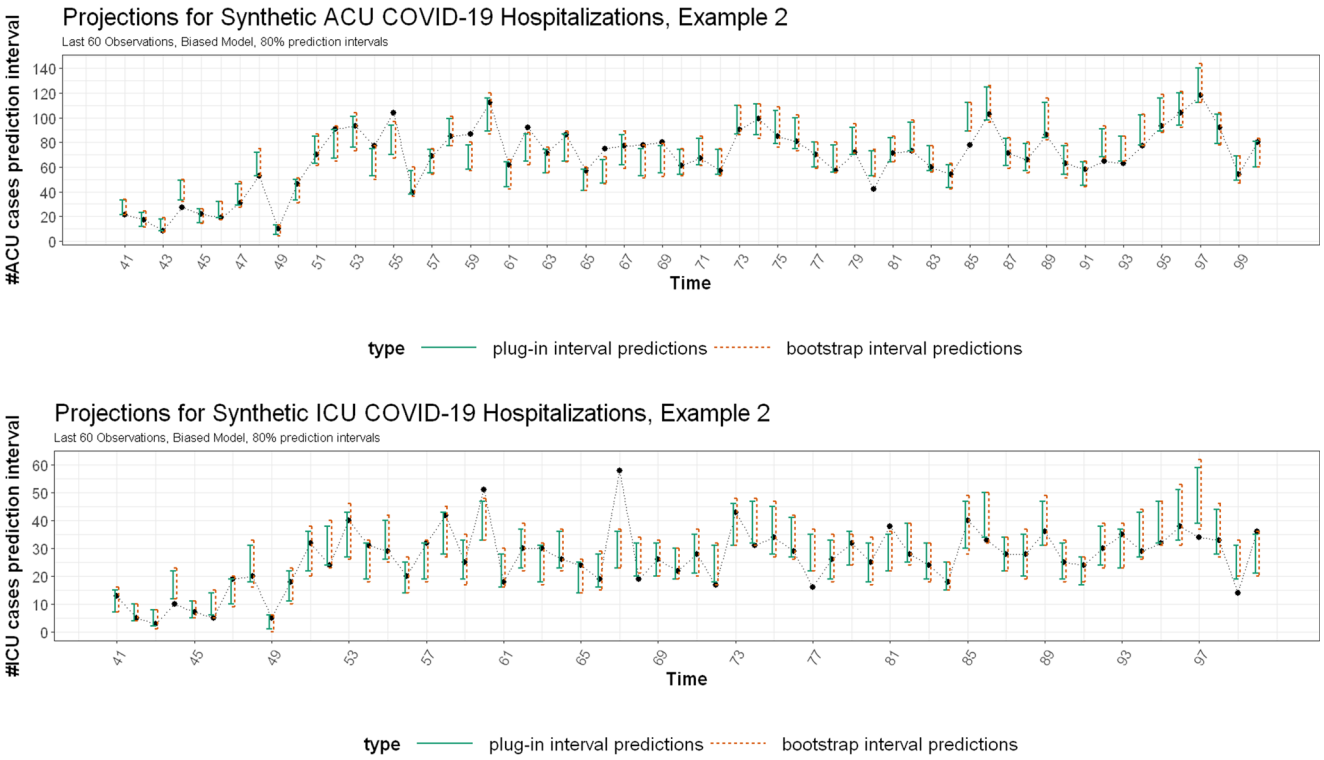


Fig. 27 Projections on synthetic data, Example 2, 80% prediction intervals, biased model, with black dots representing actual values

Table 9 Coverage rate of 80% prediction intervals, synthetic data, Example 2

Model	Plug-in, ACU	Bootstrap, ACU	Plug-in, ICU	Bootstrap, ICU
Perfect Model	78%	85%	82%	82%
Unbiased Model	77%	80%	72%	75%
Biased Model	77%	83%	78%	87%

Funding All authors report no support from external funding.

Availability of data and material County level data sources are openly available at [14] and [15]. Due to the nature of this research, the academic medical center participants of this study did not agree for their data to be shared publicly, so supporting data is not available.

Code Availability Due to the nature of this research, authors did not agree for their codes to be shared publicly.

Declarations

Conflict of Interests All authors declare no conflicts of interest with the research or writing of this paper.

References

1. Weissman GE, Crane-Droesch A, Chivers C, Luong T, Hanish A, Levy MZ, Lubken J, Becker M, Draugelis ME, Anesi GL, et al. (2020) Locally informed simulation to predict hospital capacity needs during the COVID-19 pandemic. Annals of Internal Medicine
2. Negopdiev D, Collaborative COVIDSurg, Hoste E (2020) Elective surgery cancellations due to the COVID-19 pandemic: global predictive modelling to inform surgical recovery plans. Br J Surg 107(11):1440–1449
3. Livingston E, Bucher K (2020) Coronavirus disease 2019 (COVID-19) in Italy. JAMA 323(14):1335–1335
4. Khullar D, Bond AM, Schpero WL (2020) Covid-19 and the financial health of us hospitals. JAMA 323(21):2127–2128
5. CovidActNow (2020) Covid ActNow. https://covidactnow.org/?s=1279305
6. GLEAM (2020) GLEAM Project. https://covid19.gleamproject.org/#about
7. Lemaitre JC, Grantz KH, Kaminsky J, Meredith HR, Truelove SA, Lauer SA, Keegan LT, Shah S, Wills J, Kaminsky K, et al. (2020) A scenario modeling pipeline for COVID-19 emergency planning. medRxiv
8. Altieri N, Barter RL, Duncan J, Dwivedi R, Kumbier K, Li X, Netzorg R, Park B, Singh C, Tan YS, et al. (2020) Curating a COVID-19 data repository and forecasting county-level death counts in the united states. arXiv:2005.07882
9. Ferstad JO, Gu AJ, Lee RY, Thapa I, Shin AY, Salomon JA, Glynn P, Shah NH, Milstein A, Schulman K, et al. (2020) A model to forecast regional demand for COVID-19 related hospital beds. medRxiv
10. Ray EL, Wattanachit N, Niemi J, Kanji AH, House K, Cramer EY, Bracher J, Zheng A, Yamana TK, Xiong X, Woody S, Wang Y, Wang L, Walraven RL, Tomar V, Sherratt K, Sheldon D, Reiner RC, Prakash BA, Osthus D, Li ML, Lee EC, Koyluoglu U, Keskinocak P, Gu Y, Gu Q, George GE, España G, Corsetti S, Chhatwal J, Cavany S, Biegel H, Ben-Nun M, Walker J, Slayton R, Lopez V, Biggerstaff M, Johansson MA, Reich NG (2020) Ensemble forecasts of coronavirus disease 2019 (COVID-19) in the u.s. medRxiv. https://doi.org/10.1101/2020.08.19.20177493. https://www.medrxiv.org/content/early/2020/08/22/2020.08.19.20177493, https://www.medrxiv.org/content/early/2020/08/22/2020.08.19.20177493.full.pdf
11. Kunst AL, Peralta Y, Reitsma M, Andrews JMCA, Chin L, Claypool A, Covarrubias HB, Daniels A, Fernandez M, Fung H, et al. (2020) Stanford-CIDE coronavirus simulation model (sc-cosmo)–technical description document, version 2.0
12. Pei S, Shaman J (2020) Initial simulation of SARS-CoV-2 spread and intervention effects in the continental us. medRxiv
13. Arik SO, Li C-L, Yoon J, Sinha R, Epshteyn A, Le LT, Menon V, Singh S, Zhang L, Yoder N, et al. (2020) Interpretable sequence learning for COVID-19 forecasting. arXiv:2008.00646
14. CDC (2020) COVID-19 Forecasts: Hospitalizations. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/hospitalizations-forecasts.html
15. CalCAT (2020) CalCAT Project. https://calcat.covid19.ca.gov/cacovidmodels/
16. Reich NG, Brooks LC, Fox SJ, Kandula S, McGowan CJ, Moore E, Osthus D, Ray EL, Tushar A, Yamana TK, et al. (2019) A collaborative multiyear, multimodel assessment of seasonal influenza forecasting in the United States. Proceedings of the National Academy of Sciences 116(8):3146–3154
17. McGowan CJ, Biggerstaff M, Johansson M, Apfeldorf KM, Ben-Nun M, Brooks L, Convertino M, Erraguntla M, Farrow DC, Freeze J, et al. (2019) Collaborative efforts to forecast seasonal influenza in the United States, 2015–2016. Scientific Reports 9(1):1–13
18. Poirier C, Lavenu A, Bertaud V, Campillo-Gimenez B, Chazard E, Cuggia M, Bouzillé G (2018) Real time influenza monitoring using hospital big data in combination with machine learning methods: comparison study. JMIR public health and surveillance 4(4):e11361
19. Dugas AF, Jalalpour M, Gel Y, Levin S, Torcaso F, Igusa T, Rothman RE (2013) Influenza forecasting with Google flu trends. PloS one 8(2):e56176
20. Hussain S, Harrison R, Ayres J, Walter S, Hawker J, Wilson R, Shukur G (2005) Estimation and forecasting hospital admissions due to influenza: Planning for winter pressure. the case of the west Midlands, UK. J Appl Stat 32(3):191–205
21. Araz OM, Bentley D, Muelleman RL (2014) Using Google flu trends data in forecasting influenza-like–illness related ed visits in Omaha, Nebraska. The American Journal of Emergency Medicine 32(9):1016–1023
22. Boyle JR, Sparks RS, Keijzers GB, Crilly JL, Lind JF, Ryan LM (2011) Prediction and surveillance of influenza epidemics. Medical Journal of Australia 194:S28–S33
23. Hartley DM, Giannini CM, Wilson S, Frieder O, Margolis PA, Kotagal UR, White DL, Connelly BL, Wheeler DS, Tadesse DG, et al. (2017) Coughing, sneezing, and aching online: Twitter and the volume of influenza-like illness in a pediatric hospital. PLoS One 12(7):e0182008
24. Akaike H (1969) Fitting autoregressive models for prediction. Annals of the Institute of Statistical Mathematics 21(1):243–247
25. Weron R, Misiorek A (2008) Forecasting spot electricity prices: A comparison of parametric and semiparametric time series models. International Journal of Forecasting 24(4):744–763
26. Rabiner L, Juang B (1986) An introduction to hidden Markov models. IEEE ASSP Magazine 3(1):4–16
27. Rabiner LR (1989) A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 77(2):257–286
28. Lanchantin P, Pieczynski W (2005) Unsupervised restoration of hidden nonstationary Markov chains using evidential priors. IEEE Transactions on Signal processing 53(8):3091–3098
29. Qi M, Zhang GP (2001) An investigation of model selection criteria for neural network time series forecasting. Eur J Oper Res 132(3):666–680
30. Tealab A (2018) Time series forecasting using artificial neural networks methodologies: A systematic review. Future Computing and Informatics Journal 3(2):334–340
31. Kermack WO, McKendrick AG (1991) Contributions to the mathematical theory of epidemics–I. Bulletin of Mathematical Biology 53(1-2):33–55

32. Li MY, Muldowney JS (1995) Global stability for the SEIR model in epidemiology. Mathematical Biosciences 125(2):155–164
33. Hamrock E, Paige K, Parks J, Scheulen J, Levin S (2013) Discrete event simulation for healthcare organizations: a tool for decision making. J Healthcare Management 58(2):110–124
34. Hung GR, Whitehouse SR, O'Neill C, Gray AP, Kissoon N (2007) Computer modeling of patient flow in a pediatric emergency department using discrete event simulation. Pediatric Emergency Care 23(1):5–10
35. Wood RM, McWilliams CJ, Thomas MJ, Bourdeaux CP, Vasilakis C (2020) COVID-19 scenario modelling for the mitigation of capacity-dependent deaths in intensive care. Health Care Management Science 23(3):315–324
36. Daley DJ, Vere-Jones D (2003) An introduction to the theory of point processes, volume 1: Elementary theory and methods. Springer-Verlag New York Berlin Heidelberg
37. Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC Press, Cleveland
38. Anderson TW (1971) The statistical analysis of time series. Wiley Online Library, Hoboken
39. Bartlett MS (1956) Deterministic and stochastic models for recurrent epidemics. In: Proceedings of the third Berkeley symposium on mathematical statistics and probability, vol 4, p 109
40. Allen Linda JS (1994) Some discrete-time SI, SIR, and SIS epidemic models. Math Biosci 124(1):83–105
41. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton