

## RESEARCH ARTICLE

## Real-time forecasting of COVID-19-related hospital strain in France using a non-Markovian mechanistic model

Alexander Massey<sup>1</sup>\*, Corentin Boennec<sup>2</sup>, Claudia Ximena Restrepo-Ortiz<sup>3</sup>, Christophe Blanchet<sup>4</sup>, Samuel Alizon<sup>1,5,‡</sup>, Mircea T. Sofonea<sup>1,6,7,‡</sup>

**1** Infectious Diseases and Vectors: Ecology, Genetics, Evolution and Control (MIVEGEC), Université de Montpellier, National Centre for Scientific Research (CNRS), French National Research Institute for Sustainable Development (IRD), Montpellier, France, **2** Laboratoire Plasma et Conversion d'Energie (LAPLACE), National Centre for Scientific Research (CNRS), Institut National Polytechnique de Toulouse (Toulouse INP), Université Toulouse 3-Paul Sabatier, Toulouse, France, **3** MARine Biodiversity, Exploitation & Conservation (MARBEC), Université de Montpellier, National Centre for Scientific Research (CNRS), French National Institute for Ocean Science and Technology (Ifremer), French National Research Institute for Sustainable Development (IRD), Montpellier, France, **4** Institut Français de Bioinformatique, IFB-core UAR 3601, National Centre for Scientific Research (CNRS), Évry, France, **5** Center for Interdisciplinary Research in Biology (CIRB), Collège de France, National Centre for Scientific Research (CNRS), National Institute of Health and Medical Research (Inserm), Université Paris Sciences et Lettres, Paris, France, **6** Pathogenesis and Control of Chronic and Emerging Infections (PCCEI), Université de Montpellier, National Institute of Health and Medical Research (Inserm), Établissement Français du Sang (EFS), Université des Antilles, Montpellier, France, **7** Centre Hospitalier Universitaire de Nîmes (CHU de Nîmes), Nîmes, France

\* These authors contributed equally to this work.

‡ These senior authors also contributed equally to this work.

\* [stats.massey@gmail.com](mailto:stats.massey@gmail.com)

## OPEN ACCESS

**Citation:** Massey A, Boennec C, Restrepo-Ortiz CX, Blanchet C, Alizon S, Sofonea MT (2024) Real-time forecasting of COVID-19-related hospital strain in France using a non-Markovian mechanistic model. *PLoS Comput Biol* 20(5): e1012124. <https://doi.org/10.1371/journal.pcbi.1012124>

**Editor:** Virginia E. Pitzer, Yale School of Public Health, UNITED STATES

**Received:** July 4, 2023

**Accepted:** May 1, 2024

**Published:** May 17, 2024

**Peer Review History:** PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pcbi.1012124>

**Copyright:** © 2024 Massey et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The model in this manuscript is based on publicly available data provided by Santé Publique France (<https://www.data.gouv.fr/fr/datasets/>). The scripts and data

## Abstract

Projects such as the European Covid-19 Forecast Hub publish forecasts on the national level for new deaths, new cases, and hospital admissions, but not direct measurements of hospital strain like critical care bed occupancy at the sub-national level, which is of particular interest to health professionals for planning purposes. We present a sub-national French framework for forecasting hospital strain based on a non-Markovian compartmental model, its associated online visualisation tool and a retrospective evaluation of the real-time forecasts it provided from January to December 2021 by comparing to three baselines derived from standard statistical forecasting methods (a naive model, auto-regression, and an ensemble of exponential smoothing and ARIMA). In terms of median absolute error for forecasting critical care unit occupancy at the two-week horizon, our model only outperformed the naive baseline for 4 out of 14 geographical units and underperformed compared to the ensemble baseline for 5 of them at the 90% confidence level ( $n = 38$ ). However, for the same level at the 4 week horizon, our model was never statistically outperformed for any unit despite outperforming the baselines 10 times spanning 7 out of 14 geographical units. This implies modest forecasting utility for longer horizons which may justify the application of non-Markovian compartmental models in the context of hospital-strain surveillance for future pandemics.

used to perform the analysis and generate this manuscript are available on GitLab ([https://gitlab.in2p3.fr/ete/covidici\\_public](https://gitlab.in2p3.fr/ete/covidici_public)) and archived in Zenodo (doi:10.5281/zenodo.7641132). The companion web dashboard is hosted by France Bioinformatique (<https://cloudapps.france-bioinformatique.fr/covidici/>).

**Funding:** Centre National de la Recherche Scientifique, MODCOVID19/INSMI PaSSES project to AM; Université de Montpellier to AM and CXRO; Région Occitanie, ANR Flash PHYEPI project to CB; Université de Montpellier to MTS; Centre National de la Recherche Scientifique to SA; Agence Nationale de la Recherche, ANR-11-INBS-0013 to CB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing interests:** The authors have declared that no competing interests exist.

## Author summary

The US and European Covid-19 Forecast Hubs focus on metrics such as deaths, new cases, and hospital admissions, but do not offer measurements of hospital strain like critical care bed occupancy, which was essential for the provisioning of healthcare resources during the COVID-19 pandemic. Furthermore, forecasting support was only guaranteed on the national level leaving many countries to look elsewhere for valuable sub-national forecasts. In France statistical modelling approaches were proposed to anticipate hospital strain at the sub-national level but these were limited by a two-week forecast horizon. We present a sub-national French modelling framework and online application for anticipating hospital strain at the four-week horizon that can account for abrupt changes in key epidemiological parameters. It was the only publicly available real-time non-Markovian mechanistic model for the French epidemic when implemented in January 2021 and, to our knowledge, it still was at the time it stopped in early 2022. Further adaptations of this surveillance system can serve as an anticipation tool for hospital strain across sub-national localities to aid in the prevention of short-noticed ward closures and patient transfers.

## Introduction

The COVID-19 pandemic emphasised that policymakers need access to accurate forecasts of key epidemiological indicators to mitigate strain on hospital services and reduce preventable deaths [1, 2]. This has led to international projects tasked with creating centralised repositories for COVID-19 forecasts pertaining to the United States [3], Germany/Poland [4], and Europe [5]. Sub-national forecasts were of particular interest as policies implemented at the local level tend to outperform uniform national policies [6] and as the geographic resolution of the forecasts is a relevant factor for their effectiveness [7]. However, in contrast to its counterparts, the European COVID-19 Forecast Hub, as well as many other COVID-19 forecasting analyses [8, 9], only considers the national level. This left countries like France to rely on other sources to optimise their local to national healthcare system management.

Of particular interest is intensive care unit (ICU) occupancy, one of the most direct indicators of hospital strain, which is commonly predicted using either statistical or mechanistic models. Statistical models use correlations in previously observed data to explain model structure that can be separated from noise and extrapolated to the future. In time series analysis, they can benefit from the addition of adequate predictor variables. For example, in France [10] proposed an ensemble model approach that combined several statistical models (including machine learning) to utilise predictors identified from available epidemiological, mobility, and meteorological data to make 14-day forecasts for ICU admissions and ICU occupancy by French region. Their ensemble was effective in the short-term but had more difficulty predicting beyond the lag (typically at most two weeks) between their predictors (e.g. positive antigen testing) and the subsequent hospitalisation events.

Mechanistic models are explicitly based on a simplified version of the underlying epidemiological process [11]. The most popular is the compartmental model, which typically involves separating a population into distinct sub-populations (e.g. susceptible, infected, and “removed” individuals in the SIR model) and inferring the transition dynamics between these compartments. This can be done based on assumptions regarding the biology of the pathogen or via optimisation approaches using observed data, e.g. hospital admissions, to make inferences regarding partially or unobserved factors such as daily infections [12]. The biological

assumptions simplify the causal relationships between a pathogen's infectivity, pathogenicity, and lethality so that long-term forecasts can be produced. These models provide us with a mechanistic understanding of the epidemic process and can help to anticipate planned changes such as lockdowns or increases in vaccination coverage. A limitation of compartmental models is that they typically require large, idealised populations for best results [13] and can have lower predictive performance [14] depending on the time scale. However, as shown in [15], a non-Markovian discrete-time compartmental model may have the potential to capture the dynamics up to 5 weeks on average (although this also reflects the epidemiological relevance of the underlying assumptions made).

To facilitate COVID-19 monitoring in France, we developed *COVIDici*: a mechanistic transmission model that accurately captures the hospital and mortality dynamics from the French epidemic. Model results were publicly communicated via a web dashboard (<https://cloudapps.france-bioinformatique.fr/covidici/>), which provided real-time visualisations of the French epidemic at the national, regional, and departmental levels. *COVIDici* was updated daily using databases published by the national public health agency [16] until 2022 when it was halted after the emergence of the Omicron variant [17] led to decreased interest in epidemic forecasting by French authorities, partly driven by the belief that Omicron BA.1 would represent our way out of the pandemic and the last wave [18].

Here, we briefly summarise the structure of the underlying compartmental model, the statistical procedure for the parameter inference and describe communication via the web dashboard. We present an exploratory data analysis to retrospectively evaluate *COVIDici*'s forecasts for ICU occupancy up to the four-week horizon at the regional and national levels using standard metrics for continuous variables as well as a binarised version representing ICU overload to focus on model performance in anticipating wave peaks. Standard statistical forecasting methods (auto-regression, exponential smoothing, and ARIMA) are included as baseline models. Finally, we discuss perspectives for COVID-19 epidemic modelling in the context of decreased surveillance and what this means for the surveillance of pandemics in the future.

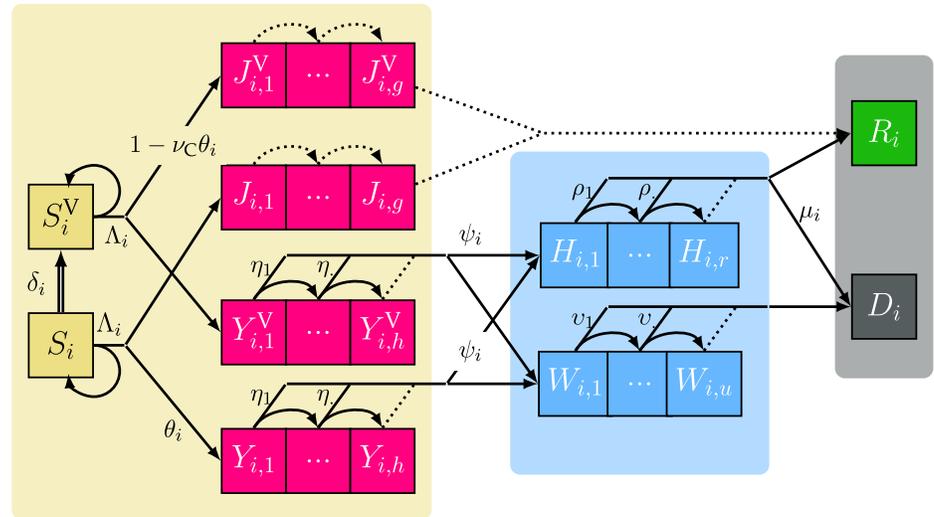
## Materials and methods

This section presents a summary of *COVIDici* and its retrospective evaluation. *COVIDici* is the sub-national extension on a pre-existing discrete-time epidemiological model (*COVIDSIM-FR*, [19]), developed in R [20], that combines the computational benefits of deterministic dynamical systems with the short-time accuracy of non-Markovian dynamics. *COVIDSIM-FR* was initially tailored to capture the dynamics of national ICU bed occupancy during the first COVID-19 wave in France (March-May 2020) [21], and *COVIDici* was further developed to pursue this objective at sub-national levels and taking into account both the vaccination campaign and viral evolution.

The scripts and data used to perform the analysis and generate this manuscript are available on GitLab ([https://gitlab.in2p3.fr/ete/covidici\\_public](https://gitlab.in2p3.fr/ete/covidici_public)) and archived in Zenodo [22]. The current section provides an overview of the main methods used in both the implementation of *COVIDici*, while further technical details regarding the implementation of *COVIDici* as well as all the baseline models is accessible in [S1 Text](#). The mathematical and computational foundations of the model behind *COVIDici* are exposed in [21].

## The model

The structure of the model, shown in [Fig 1](#), describes the flows between the relevant clinical-epidemiological compartments within hospitals as well as the greater community. It includes



**Fig 1. Structure of the underlying COVID-19 epidemic discrete time model with vaccination.** The three shaded areas represent the general population or community (left), the critically-ill hospitalised patients (center) and individuals removed by either recovery or death (right). The first subscript ( $i$ ) of compartment— $g$ ,  $h$ ,  $r$ , and  $u$  are thus the maximum number of days possible to remain in a compartment represented by contiguous boxes (see [21] for their computation). Individuals in  $S$  are susceptible,  $J$  are non-critically infectious,  $Y$  are infected and will eventually require hospitalisation,  $H$  are hospitalised in a critical care bed for a long stay,  $W$  are in non-ICU beds,  $R$  are recovered and  $D$  are deceased.  $S^V$ ,  $J^V$  and  $Y^V$  are the vaccinated counterparts of  $S$ ,  $J$  and  $Y$ .  $\Lambda_i$  is the daily force of infection (a probability).  $\delta_i$  is the daily vaccination rate.  $\mu_i$ ,  $\psi_i$ , and  $\theta_i$  are transition rates between compartments, the latter being reduced by the factor  $v_C$  for vaccinated individuals. Arrows between boxes show the daily flow of individuals between compartments where dotted arrows occur with probability 1. For the sake of simplicity, only one age group is depicted here and only one of the two complementary probabilities is shown for each bifurcating transition.

<https://doi.org/10.1371/journal.pcbi.1012124.g001>

individuals that, by assumption [23], no longer contribute to the epidemic due to recovery with full immunity or death. Stratifying by age, where the index  $i$  denotes the individual age class, each box represents a group of individuals who share the same clinical kinetics and who contribute equally to the epidemic dynamics. Most susceptible individuals in  $S_i$  will pass to the non-critically ill compartment, denoted  $J_b$ , with rate proportional to  $(1 - \theta)$ . A small fraction  $\theta$  of the infected individuals, which increases with age [24], reaches the critical infection compartment,  $Y_b$ , meaning they will eventually be hospitalised either for a long stay in critical care wards,  $H_b$ , or they will be in conventional hospital care ( $W_i$ ) where they end up in the deceased compartment  $D_i$  with probability 1. Critically-ill individuals move to  $H_i$  with probability  $\psi_i$  and might die with probability  $\mu_i$ . Those who do not die enter the recovered compartment  $R_i$  where they are assumed to have perfect immunity for the rest of the simulation period.  $\eta$ ,  $\rho$  and  $\nu$  reflect the daily probabilities to exit  $Y_b$ ,  $H_i$  and  $W_i$  respectively.

Aged care facilities (i.e. nursing homes) are ignored in this model because of their differences in hospitalisation rates and epidemiological dynamics compared to the general population [25]. Importantly, individuals are stratified according to vaccination status (upon their first dose) prior to hospitalisation: the vaccinated compartments are denoted by the exponent  $^V$ .  $\delta_i$  is the daily vaccination rate for age class  $i$ , given by, or linearly extrapolated according to the trend of the previous 3 weeks, from the French vaccination database VAC-SI [26].

The most important quantity of the COVID<sub>ici</sub> model is the daily rate at which susceptible hosts become infected, namely the discrete-time force of infection (FOI)  $\Lambda_i(t)$ . For a given

national or sub-national location  $\ell$  and calendar date  $t$ , FOI is defined by

$$\Lambda_{i,\ell}(t) := \left( 1 + \left( \frac{\mathcal{R}_0}{S_\ell^\circ} C_\ell(t) \sum_{\tau=1}^{\tau_{\max}} \zeta_\tau \sum_{j \in \mathcal{A}} J_{j,\tau}(t) + Y_{j,\tau}(t) + v_T (J_{j,\tau}^Y(t) + Y_{j,\tau}^Y(t)) \right)^{-1} \right)^{-1}, \quad (1)$$

where  $\mathcal{R}_0$  is the basic reproduction number of the original (Wuhan) strain of SARS-CoV-2 in France,  $S^\circ$  is the population size in location  $\ell$ , and  $C_\ell(t)$  is the piecewise-constant transmission factor that captures all spatiotemporal changes in SARS-CoV-2 spread. Importantly, temporal variations of  $C_\ell(t)$  are likely induced by non-pharmaceutical interventions (NPIs) [27], spontaneous behavioural changes and viral evolution [28] while the spatial covariance of  $C_\ell(t)$  reflects the heterogeneity in the contact rate between departments/regions, in variant spread as well as in NPI implementation and compliance [29, 30]. The index  $\tau$  denotes the time in days since the beginning of the infection and allows us to vary host infectivity over the duration of the infection. For this,  $\zeta_\tau \in [0, 1]$  is the generation time probability mass (i.e. the relative contribution to infectivity) of the  $\tau$ -th day, for which we use the empirical serial interval distribution estimated by [31]. Finally, the  $v_T = 0.2$  ratio captures the reduction in viral transmission due to vaccination (mainly due to infection prevention [32]). The rationale of the Holling’s type II functional response of the FOI (analogous to Michaelis-Menten kinetics) is elaborated in the Appendix of [21].

### Modelling vaccination

The French national vaccination campaign started in late December 2020. To incorporate the vaccine rollout into the model, we explicitly assumed that the vaccines partially prevent infection and critical COVID-19 by reducing the probability of being critically ill upon exposure, denoted by factor  $v_C$ . We set vaccine coverage in each age class in the model using the official VAC-SI database [33].

Future vaccination rates were predicted using a linear regression for each age group trained on the previous 3 weeks of vaccination data. We assumed that vaccination begins with the older age classes and that all age classes have an arbitrary vaccine coverage threshold of 90%. If the coverage for an age class was ever over this threshold, the doses planned for this age class were redistributed to the next oldest age class.

To avoid the inflation of the number of parameters, we assumed that full vaccination only required a single dose. Another simplifying assumption is that the vaccine is instantaneously effective with an assumed permanent reduction in critical infection upon exposure of 90% ( $v_C = 0.1$ ), which is in the order of magnitude of the first real-life estimates available at the time of model development [32].

### Calculation

Parameter inference relies on a computed distance of the daily ICU admissions simulated by the model with respect to publicly reported data [34] after treatment for weekly seasonality. Let us denote the publicly produced figures for the number of ICU admissions in location  $\ell$  (whether at the departmental, regional or national level) on calendar date  $t$  by  $a_{\ell,t}$ . The weekly seasonality, being caused by systematic under-reporting on weekends and over-reporting during the beginning of the following work week, was smoothed out using 7-day rolling average followed by a Gaussian rounding (denoted by  $\lfloor \cdot \rfloor$ ). The observed data considered for the inference procedure is therefore  $a_{\ell,t}^\circ := \left\lfloor \frac{1}{7} \sum_{\tau=0}^6 a_{\ell,t-\tau} \right\rfloor$  and let  $\mathcal{T}$  be the set of calendar dates  $t$  for which  $a_{\ell,t}^\circ$  is computable.

Denoting by  $a_{\ell,t}^s(\mathbf{x})$  the number of ICU admissions in location  $\ell$  at time  $t$  simulated by the model being parametrised with the set of parameter values  $\mathbf{x}$ , the log-likelihood of  $\mathbf{x}$  given the observed data is computed as follows:

$$\log \mathcal{L}_{\ell, \mathcal{T}}(\mathbf{x}) := \log \mathbb{P}[(a_{\ell,t}^{\circ})_{t \in \mathcal{T}} | (a_{\ell,t}^s(\mathbf{x}))_{t \in \mathcal{T}}] := \sum_{t \in \mathcal{T}} a_{\ell,t}^{\circ} \log(a_{\ell,t}^s(\mathbf{x})) - \log(a_{\ell,t}^{\circ}!) - a_{\ell,t}^s(\mathbf{x}). \quad (2)$$

This definition holds if the distance between the model and the observation on a given date is seen as the random fluctuation of an integer-valued random variable. The random variable can reasonably be assumed to be Poisson-distributed. This assumption works well for small admission numbers because the population sizes of the investigated locations are large while the daily individual probability of being admitted in an ICU for COVID-19 is small. Assuming that the random fluctuations are independent across days and locations, the log-likelihood over a given set of dates  $\mathcal{T}$  can be easily computed by summing over the daily log-likelihoods.

Parameters were then inferred under the Bayesian framework by running a Markov chain Monte Carlo (MCMC) algorithm—implemented in the `BayesianTools` R package [35]—over a set of 12,000 realisations of the model. The last 2,000 iterations were used to generate the median and the equal-tailed 95% credibility intervals for the parameters as well as the 95% forecasting range for the extrapolated time series of daily ICU admissions assuming a Poisson likelihood. The reproduction number  $\mathcal{R}_0$  and the expectation and variance of the infection-to-hospitalisation delay were inferred at the nationwide level only, while all other parameters were independently fit for each sub-national administrative division. Details on the inferred parameters, their prior values and distributions are provided in Tables A, B, C, D and E in [S1 Text](#) and in [21].

We expected some of these parameters to change over time due a variety of factors including virus evolution (e.g. the increased transmissibility of the Alpha [28] and Delta [36] variants), public health interventions (e.g. lockdowns, curfews, limitations on businesses, etc.), social factors (e.g. school holidays), improvement in COVID-19 patient care, and variation in patient profiles. To account for this, we allowed for some parameters to be time-dependent by partitioning the time since the beginning of the epidemic and allowing each period to be associated with its own parameter set.

Parameter estimations were optimised based on the daily COVID-19-related critical care admissions from the COVID-19 hospital activity database (SI-VIC) [16]. In France, critically ill patients can be hospitalised either in intensive care units, continuous care units, or acute care units, with the three forming the critical care capacities [37]. For simplicity here, ICU refers to the wider category of critical care beds, as provided by SI-VIC. Furthermore, we assume the age distribution between localities to be fixed and based on the official demographics data [38].

## Communication

An automated cluster computing workflow refit the `COVIDici` model using daily updates of hospital, vaccination and testing data downloaded from the SI-VIC database, allowing a Shiny web application (see link in Introduction or [22] for source code) to communicate real-time results to the public. The original 2021 production version permitted users to visualise the combined past and future model fit by national, regional or departmental administrative unit for multiple epidemiological parameters, including ICU admissions, ICU occupancy, mortality (cumulative and daily), temporal reproductive number ( $R_t$ ), infections (cumulative, daily and current), vaccination coverage and incidence for positive tests.

In 2022, a post-mortem version of the interface was deployed to allow for retrospective inspection of past forecasts with respect to a historical reference date. This version allows visualisation of all forecasts occurring prior to the reference date and includes basic evaluation metrics based on ICU overload (i.e. binarised ICU occupancy) and a colour-coded heatmap of hospital strain for varying forecast lengths and arbitrary saturation thresholds.

### Retrospective evaluation

Our assessment is based on original forecasts made by COVIDici between January 30, 2021 and December 2, 2021 (i.e. the first detected Omicron case in France), taken on a weekly basis to match with evaluation frameworks of the European and US Covid-19 Forecast Hubs. We focus here on ICU occupancy and only consider the regional and national levels while emphasising that many of the results are equally as valid on the departmental level, especially when they contain major urban areas. The following baseline models were evaluated using a rolling forecasting origin [39] starting on August 2, 2020:

- **ETS+ARIMA** is an ensemble of an ARIMA and an exponential smoothing (ETS) model fit using automated defaults in the `fable` package in R. It uses a log transformation of the rolling 7-day average of the ICU occupancy using only the data available to COVIDici to make its original forecast for that reference date.
- **AR-Lasso** is an auto-regressive (AR) machine learning type model that does not require stationarity. Lagged values from the previous 21 days are selected using the Least Absolute Shrinkage and Selection Operator (LASSO) [40] tuned using a 14-day time-series cross-validation as implemented in [41] to prevent data leakage and reduce overfitting. Prediction intervals were calculated using a bootstrap of the one-step ahead residuals from the training fit.
- **Naive** is a special case of an AR-1 implemented using the `fable` package. The point forecast is simply the last observed value and is optimal if the time series is a random walk [39].

### Standard metrics

We define “standard metrics” as those recommended by the US and European COVID-19 Forecast Hubs using the same 23 quantiles of the forecast distribution. Thus, point forecasts (based on the median) are evaluated with the absolute error (AE), individual prediction levels with the empirical coverage rate (ECR), and the forecast distributions with the weighted interval score (WIS). The WIS is a proper scoring rule that generalises the absolute error and gives penalties for interval spread as well as for over- and under-prediction. Formally speaking, for the true value of the target variable  $y$ , we have a forecast distribution  $F$  with median  $m$  that contains a set of  $K$  prediction intervals whose respective upper ( $u$ ) and lower ( $l$ ) limits are the  $\frac{\alpha}{2}$  and  $1 - \frac{\alpha}{2}$  quantiles of the  $F$ . WIS is defined as the following weighted sum:

$$\text{WIS} = \frac{1}{K + 0.5} \left( 0.5 \cdot |y - m| + \sum_{k=1}^K \frac{\alpha_k}{2} \cdot \text{IS}_{\alpha_k}(F, y) \right) \quad (3)$$

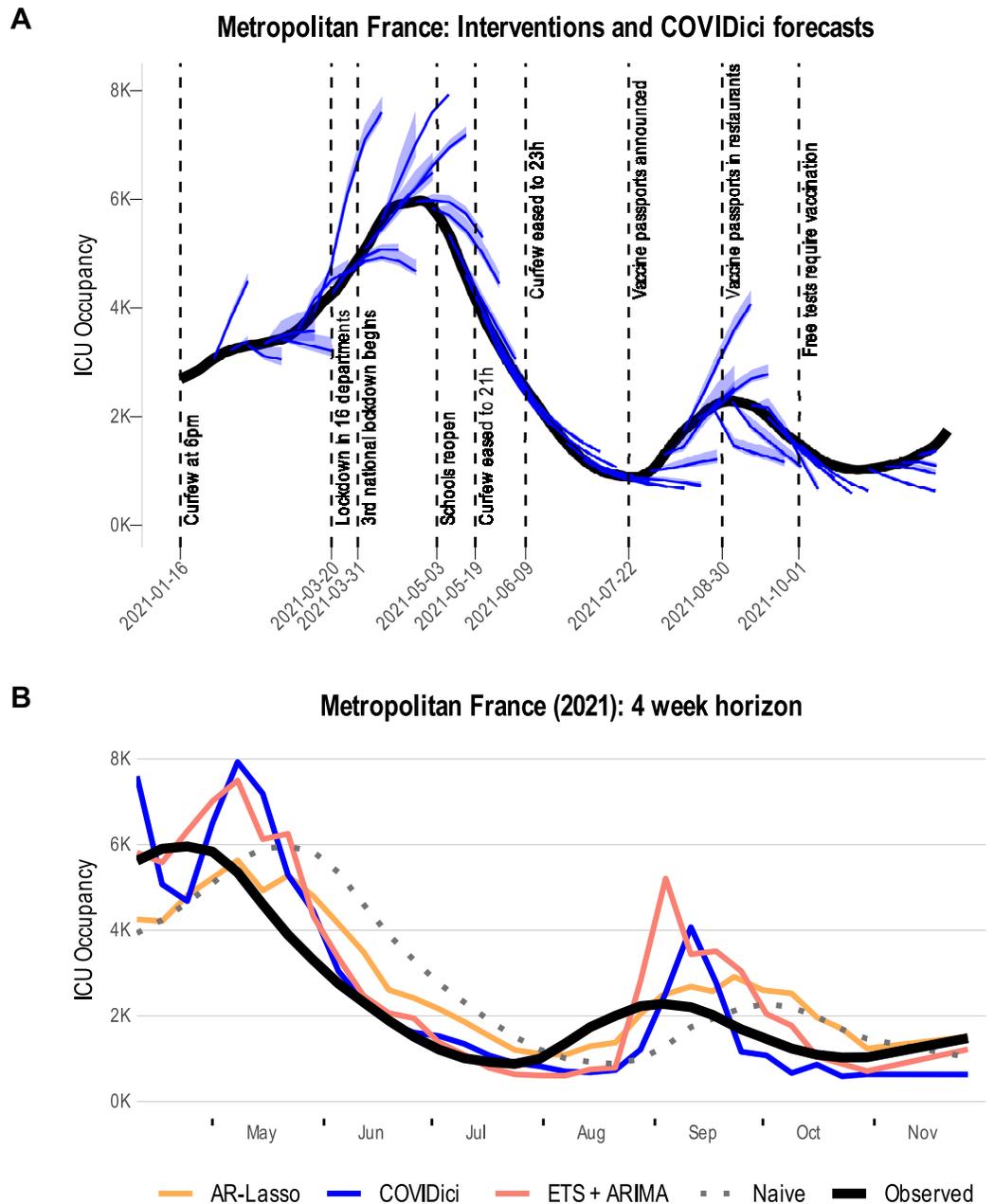
where for a single interval  $k$ , the interval score is computed as

$$\text{IS}_{\alpha_k}(F, y) = (u - l) + \frac{2}{\alpha_k} \cdot (l - y) \cdot \mathbb{1}(y \leq l) + \frac{2}{\alpha_k} \cdot (y - u) \cdot \mathbb{1}(y \geq u) \quad (4)$$

with  $\mathbb{1}$  being an indicator function and  $\alpha_k$  the nominal coverage of interval  $k$ . Note that if only the median is included (i.e.  $K = 0$ ) then the WIS simplifies to the AE. Furthermore, as the

number of equally spaced intervals increases, the WIS converges to the Continuous Ranked Probability Score. We refer readers to [42] for a deeper technical explanation.

All three standard metrics (AE, ECR, WIS) were calculated using the `scoringutils` package [43]. We used the package’s default summary function (i.e. the mean) when aggregating only over geographic units. However, we use the median when aggregating over time which is not uncommon in the forecasting literature for COVID-19 [44–46] as it is more robust to the abnormally large errors that are common during the peaks of epidemiological waves. These exaggerated errors are clearly visible for COVIDici in Fig 2A and for ETS+ARIMA in Fig 2B



**Fig 2. Qualitative inspection of COVIDici forecasts.** A) Overlay of national forecasts of ICU occupancy produced by COVIDici with list of governmental interventions issued (national level only). B) All forecasters at the 4-week horizon plotted with the observed ICU occupancy for metropolitan France.

<https://doi.org/10.1371/journal.pcbi.1012124.g002>

where its 4-week forecast horizon for metropolitan France more than doubled the observed ICU occupancy in early September 2021.

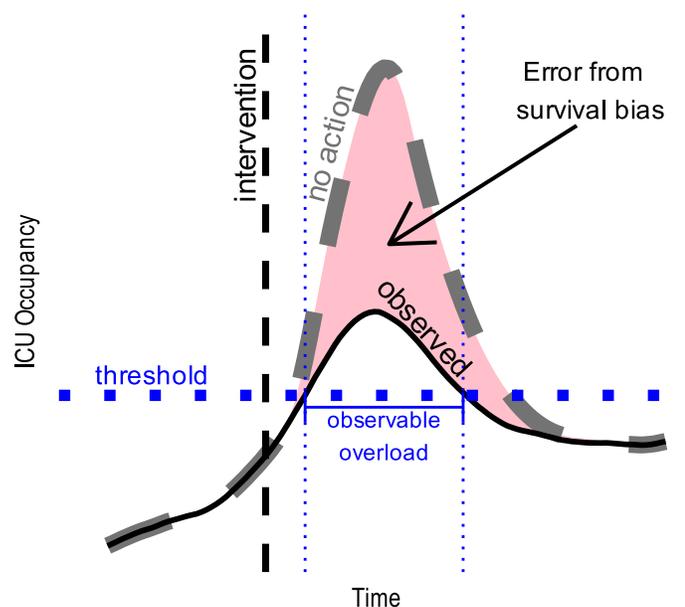
Summarising AE and WIS with the median have the drawback that it is more likely to reflect forecaster performance between waves rather than its ability to anticipate peaks, which is arguably more important depending on the forecaster's objective. Furthermore, AE and WIS tend to harshly punish large errors during wave peaks which can at least be partially explained by a survivor bias that occurs every time a public health policy is implemented (see Fig 2A for non-exhaustive list of national interventions in France), as well as spontaneous behavioural change. As artistically illustrated by Fig 3, this bias (i.e. the shaded area between the curves) corresponds to the difference between what would happen in absence of intervention (dashed curve) and what eventually is observed (the solid curve). While the magnitude of this bias at the peaks is counterfactual and subject to debate, several studies have indicated that even mild non-pharmaceutical interventions can have similar effects on curbing the spread of the virus compared to more severe ones [47, 48].

### Binary metrics

To more fairly evaluate performance during wave peaks, we consider ICU overload (i.e. binarised ICU occupancy), which we expect to be more robust against over-predictions. This requires introducing arbitrary capacity thresholds which we define as the percentage of the ICU occupancy observed in the geographical unit in the first wave in 2020. For point forecasts, we consider the proportion of incorrect forecasts of an outcome given that outcome was observed. Following the convention that lower scores are better, we define:

$$1 - \text{Sensitivity} = \frac{\# \text{ of incorrect forecasts of overload}}{\# \text{ of observed overloads}} \quad (5)$$

$$1 - \text{Specificity} = \frac{\# \text{ of incorrect forecasts of underload}}{\# \text{ of observed underloads}} \quad (6)$$



**Fig 3. Artistic representation of the survivor bias that may occur after a restrictive governmental intervention.** The dashed curve is the counterfactual ICU occupancy that would have occurred if the intervention did not happen. The solid curve is the ICU occupancy that we observed because the intervention did occur.

<https://doi.org/10.1371/journal.pcbi.1012124.g003>

For forecast distributions of ICU overload, we use the Brier score, which is the mean squared error of the binary overload outcome (i.e. 0 or 1) and the mass of the prediction interval above the arbitrary threshold. Formally, this is defined as:

$$\text{Brier score} = \frac{1}{N} \sum_{n=1}^N (f_n - y_n)^2 \quad (7)$$

where  $f_n$  is the predicted probability of overload event  $y_n \in \{0, 1\}$  with  $n = 1, \dots, N$  denoting all the events in the scope of the evaluation. The predicted probability  $f_n$  can be approximately calculated as

$$f_n \approx 1 - \min\{q \mid q \geq \text{threshold}\} \quad (8)$$

where  $q$  is the collection of quantiles of the forecast distribution and the threshold is fixed and arbitrary. We present binary metrics separately for periods of observed overload and underload because anticipation of overload is widely considered more important in a hospitalisation surveillance system.

### Statistical comparisons of forecast metrics

Many statistical approaches for COVID-19 forecast comparison have been proposed that focus primarily on producing p-values for hypothesis testing. Some notable examples include Diebold-Mariano (DM) tests [49], permutation tests (implemented in `scoringutils` for mean only), Wilcoxon signed-rank test (also implemented in `scoringutils`) and Mood's Median test [44]. While all these approaches can be quite suitable for many forecasting situations, the latter is not applicable to comparing groups paired by target date and the former three may be negatively affected when comparing two forecast error distributions with different shapes [50–52]. Recall that abnormally large errors in `COVIDici` and `ETS+ARIMA` at the wave peaks indeed create much heavier tails in their respective forecast error distributions compared to the Naive and AR-Lasso models (for example see Fig B in [S1 Text](#)). Furthermore, all these tests ignore potential serial dependencies when aggregating across geographic units while only the DM test accounts for temporal auto-correlation.

The goal of the statistical evaluation presented here is to retrospectively identify significant patterns in the forecast performance of `COVIDici`, and, as such, is exploratory in nature not confirmatory [53, 54]. Given the multitude of potential comparisons to be made across varying geographical units, forecast horizons and time, we present statistical comparisons between two models by focusing primarily on confidence intervals (CIs) that can be integrated into interpretable graphics. In particular, we implement CIs using a non-parametric bootstrap with 10,000 replicates for the ratio of the aforementioned summary metrics. If the summary metric aggregates over only time then forecast dates are independently resampled with replacement and if aggregation occurs over both space and time then both geographic units and forecast dates are simultaneously resampled. By consequence, potential serial dependences across space and time are ignored.

To account for skewness in the distribution of the bootstrapped test statistic we apply bias-corrected and accelerated (BCa) CIs [55]. The main advantage of the BCa bootstrap approach is that it is general enough to be applied to every metric regardless of potential skewness in one or both of the forecast error distributions. This adjustment also ensures that the CIs are second-order accurate and transformation invariant, which is a convenient property when presenting plots on the log scale. The main disadvantage is that we resample independently in a manner that ignores potential spatial and temporal dependency. The acceleration parameter that accounts for skewness is calculated using a finite-sample jackknife. A formal explanation

of the confidence interval implementation is notationally tedious so it is only provided in [S1 Text](#). However, convenient practical implementation is done using the `npctest` package [56] to produce CIs at the 95% level.

P-values are calculated post hoc by CI inversion [57, 58], which simply means that the p-value for an arbitrary value (i.e. 1 for a ratio comparison and 0 for a log-ratio comparison) is defined as the smallest  $\alpha$  such that the corresponding  $1 - \alpha$  CI does not contain that value. Being an exploratory data analysis, we designate important levels for p-values at  $\alpha = 0.001$ , 0.01, 0.05 and 0.1, but emphasise to the reader that these are intended to be interpreted in conjunction with their associated  $1 - \alpha$  CIs rather than being formal hypothesis tests where it is preferable to resample under the assumption that the null hypothesis is true. Practically speaking, the null hypothesis is that the ratio of summary metric statistics is 1 (0 on the log scale) and is rejected when this value not is contained in the CI.

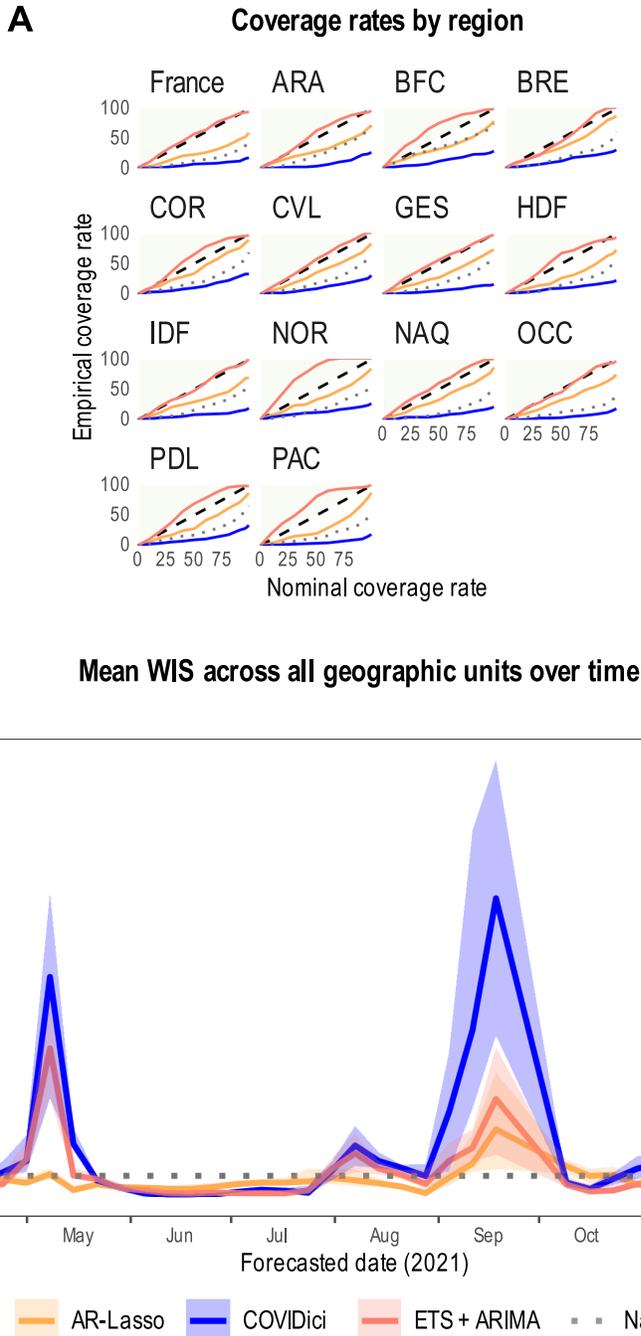
## Results

To qualitatively inspect the forecasts produced by `COVIDici` relative to the observed (smoothed) ICU occupancy time series, we first refer to [Fig 2A](#). It is noteworthy that `COVIDici` performs better on the trailing side of each wave than the leading side, a phenomenon that is also apparent on the regional and departmental level (see Post-mortem version of the Shiny app linked in the introduction). It is not completely clear why this is case but this possibly indicates more uncertainty in changes in social behaviour and effectiveness of governmental interventions as the number of infections surges. Furthermore, the turning points at the wave peaks and wave troughs (i.e. local maxima/minima) are characterized by large misses. To a certain extent this is to be expected since `COVIDici` is informed by changes in ICU admissions which may take approximately 2 weeks to manifest following a surge or decline in new infections.

[Fig 2B](#) visualises the national forecasts at the 4-week horizon for `COVIDici` and the 3 baseline models. The ensemble of ARIMA+ETS exhibits a similar behaviour to `COVIDici` where it greatly over-predicts at the top of the waves and shows the same delay in detecting the beginning of a new one. AR-Lasso and the Naive model on the other hand appear to predict a right-shifted version of the observed ICU occupancy time series. For the Naive model this shift is exact and equal to the length of the forecast horizon.

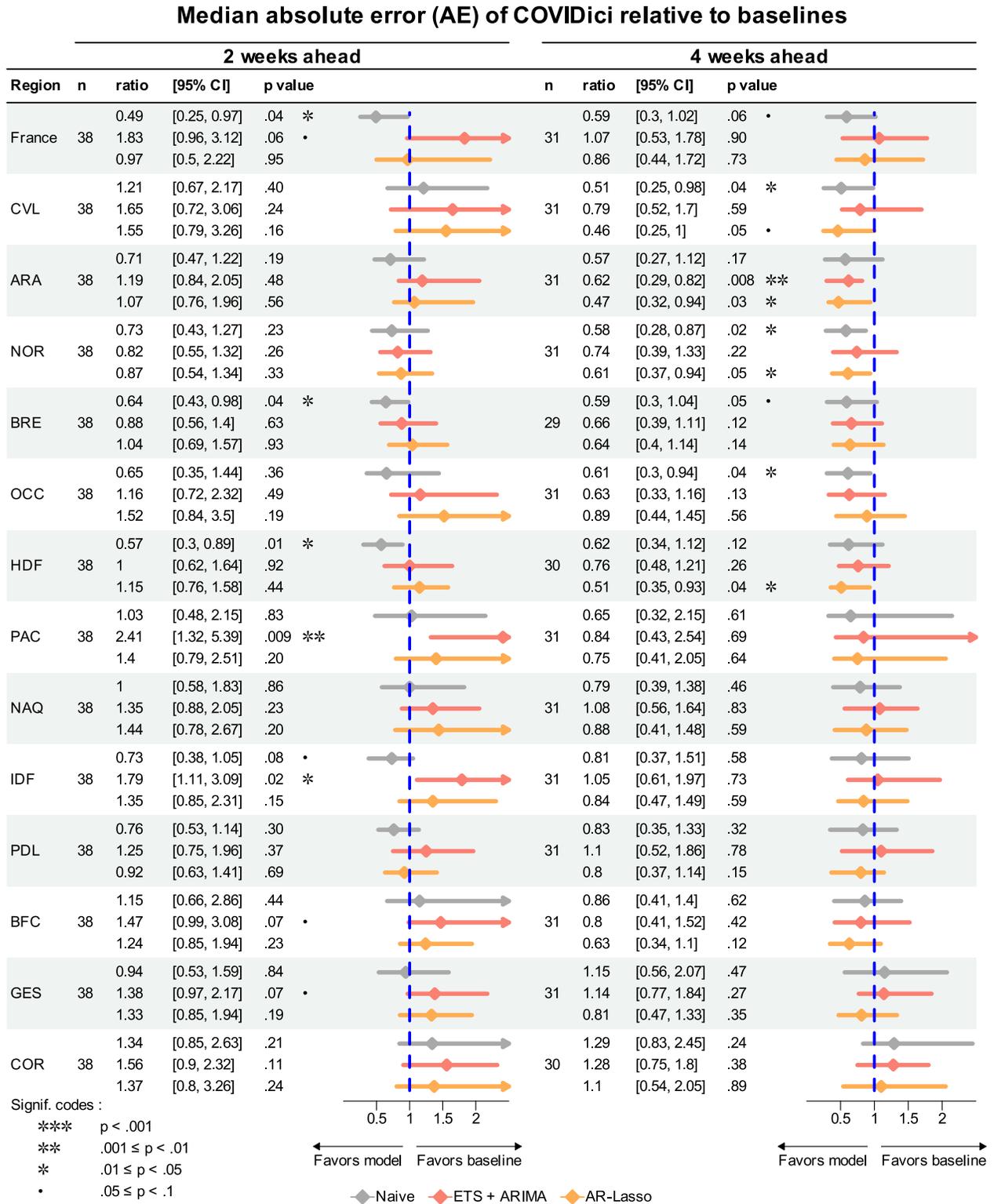
Standard metrics for ICU occupancy are contained in [Fig 4](#). [Fig 4A](#) shows the overall empirical coverage rates across all geographic units, forecast dates and forecast horizons where the dashed line is optimal. Only ETS+ARIMA appears to exhibit a reasonably calibrated forecast distribution, followed by AR-Lasso and the Naive model. The worst calibration by far are the prediction intervals for `COVIDici` which are almost horizontal in [Fig 4A](#), indicating that the prediction intervals are far too narrow. This occurred because the forecast distribution implemented by `COVIDici` is merely the credible interval for a model parameter, i.e. ICU occupancy, rather than a proper prediction interval describing the uncertainty impacting a future observation. As a result, `COVIDici`'s performance relative to all metrics for forecast distributions will be negatively affected as can be visualised in [Fig 4B](#) which shows the mean WIS (scaled by the naive model) across all geographic units over time. Large misses near wave peaks are evident for `COVIDici` and ETS+ARIMA but not AR-Lasso, which appears to be more consistent.

[Fig 5](#) shows a forest plot by region of the median AE of `COVIDici` relative to the baseline models. At the two-week horizon `COVIDici` outperformed the Naive baseline in metropolitan France ( $p = .06$ ), Bretagne ( $p = .04$ ), Hauts-de-France ( $p = .01$ ) and Île-de-France ( $p = .08$ ), but actually underperformed the ensemble ETS+ARIMA baseline in metropolitan France



**Fig 4. Standard evaluation metrics for ICU occupancy.** A) Empirical coverage rate across all geographic units and forecast horizons (dashed line is optimal). Auvergne-Rhône-Alpes = ARA, Bourgogne-Franche-Comté = BFC, Bretagne = BRE, Centre-Val de Loire = CVL, Corse = COR, Grand Est = GES, Hauts-de-France = HDF, Île-de-France = IDF, Normandie = NOR, Nouvelle-Aquitaine = NAQ, Occitanie = OCC, Pays de la Loire = PDL, Provence-Alpes-Côte d’Azur = PAC. B) Mean weighted interval score (WIS) of all geographic units over time for the four-week forecast horizon scaled by the Naive model. Shaded areas represent 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals with 10000 replicates per target date.

<https://doi.org/10.1371/journal.pcbi.1012124.g004>



**Fig 5. Median absolute error (AE) for COVIDici relative to other models.** Ratio = median AE for COVIDici / median AE for baseline model. 95% confidence interval (CI) is the bias-corrected and accelerated (BCa) bootstrap confidence interval generated by a nonparametric bootstrap with 10000 replicates. The p value is the smallest alpha such that 1 is not contained in the corresponding  $1 - \alpha$  CI. See Fig 4 for region code definitions.

<https://doi.org/10.1371/journal.pcbi.1012124.g005>

( $p = .06$ ), Provence-Alpes-Côte d'Azur ( $p = .009$ ), Île-de-France ( $p = .02$ ), Bourgogne-Franche-Comté ( $p = .07$ ) and Grand Est ( $p = .07$ ). However, at the four-week horizon COVID<sub>i</sub>ci was never statistically outperformed by any baseline model yet did better than at least one of the baselines in 7 of the 14 considered geographic units: compared the Naive model in metropolitan France ( $p = .06$ ), Centre-Val de Loire ( $p = .04$ ), Normandie ( $p = .02$ ), Bretagne ( $p = .05$ ) and Occitanie ( $p = .04$ ); compared to ETS+ARIMA in Auvergne-Rhône-Alpes ( $p = .008$ ); and compared to AR-Lasso in Centre-Val de Loire ( $p = .05$ ), Auvergne-Rhône-Alpes ( $p = .03$ ), Normandie ( $p = .05$ ) and Hauts-de-France ( $p = .04$ ). This reflects a modest but consistent improvement in forecast performance at the longer forecast horizon. Furthermore, it is unlikely that these results at the four-week horizon occurred simply due to a multiple testing problem since if the null hypothesis is true then the p-values should be uniformly distributed and all significant tests favored COVID<sub>i</sub>ci.

Fig 6 shows a similar forest plot except using median WIS as the summary metric. The overly-narrow prediction intervals clearly degraded COVID<sub>i</sub>ci's performance here as it never outperformed any of the baselines at the two-week horizon and only performed better compared to ETS+ARIMA in Normandie ( $p < .001$ ). It was also outperformed by at least one of the baselines in 10 of 14 considered geographic units.

Binarised metrics for ICU overload are shown in Fig 7. The left column shows the raw summary metrics without CIs while the right column shows the performances of the baselines relative to COVID<sub>i</sub>ci including their respective bootstrapped CIs. The relative metrics are on the log scale for more visual clarity and are interpreted to be statistically significant for a given capacity threshold when the CI of the baseline model does not contain zero (i.e. the horizontal blue line).

COVID<sub>i</sub>ci performed poorly at the two-week horizon for nearly all metrics presented here except relative to the Naive model when overload was observed at capacity thresholds greater than approximately 0.75. On the other hand, on the four-week horizon, COVID<sub>i</sub>ci outperforms AR-Lasso and the Naive model for most thresholds considered when overload was observed. For the % of incorrect forecasts at the longer horizon we fail to detect any statistical difference between COVID<sub>i</sub>ci and ETS+ARIMA for most capacity thresholds. For Brier score, COVID<sub>i</sub>ci's performance was somewhat degraded by its overly-narrow prediction intervals just as was the case for other aforementioned distribution metrics.

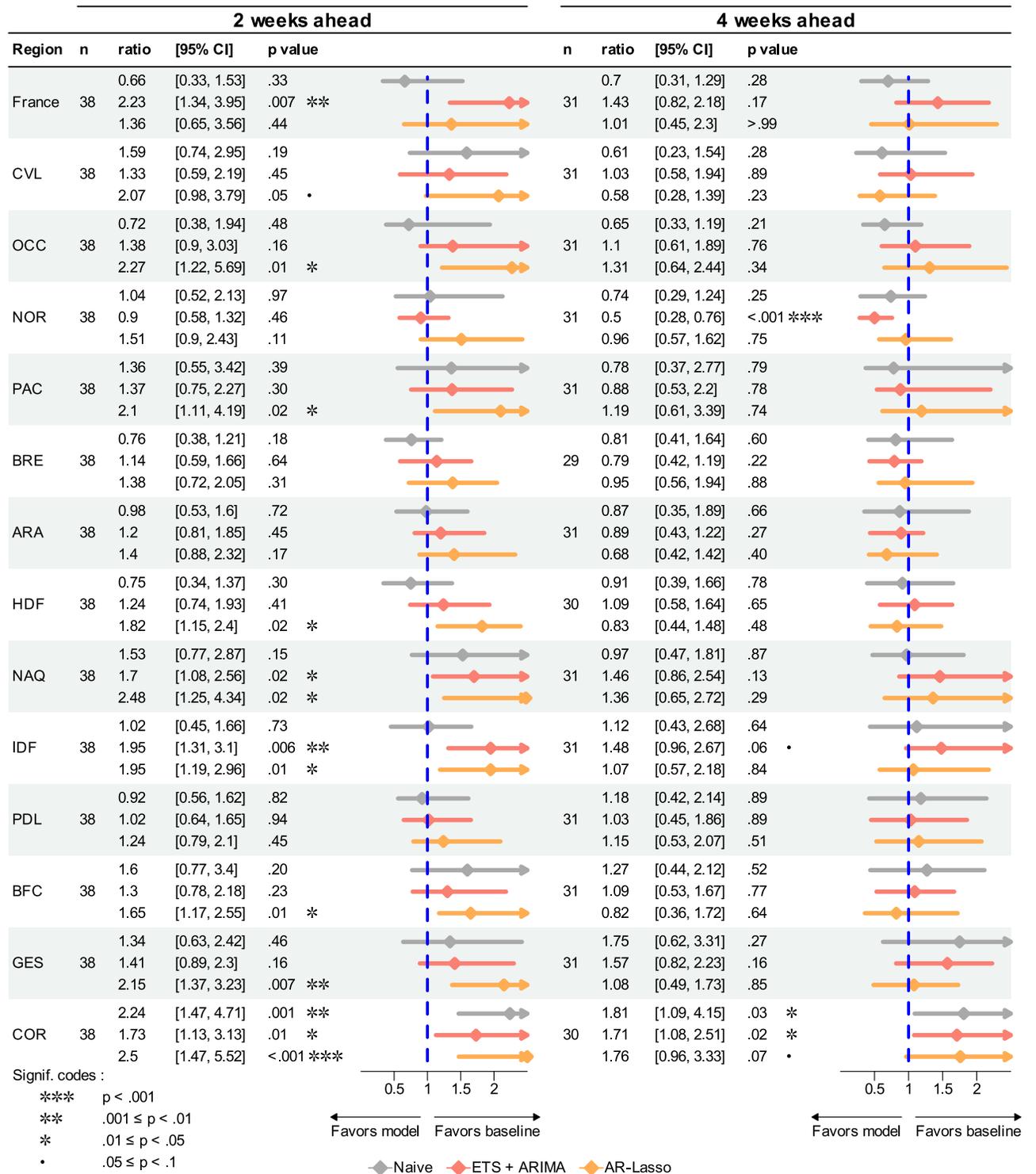
To summarise, COVID<sub>i</sub>ci tended to exhibit improved forecasting performance at the four-week horizon than the two-week horizon relative to the considered baseline models. However, its prediction intervals were far too narrow, which led to poorer performance relative to metrics for forecast distributions. In terms of median AE it either performed comparably or better than all baselines at the four-week horizon for all considered sub-national geographic units. In terms of correctly predicting ICU overload, COVID<sub>i</sub>ci was comparable to the ETS+ARIMA ensemble model at the four-week horizon and consistently outperformed the Naive and AR-Lasso baselines for most of the thresholds considered.

## Discussion

### Modelling and forecasting

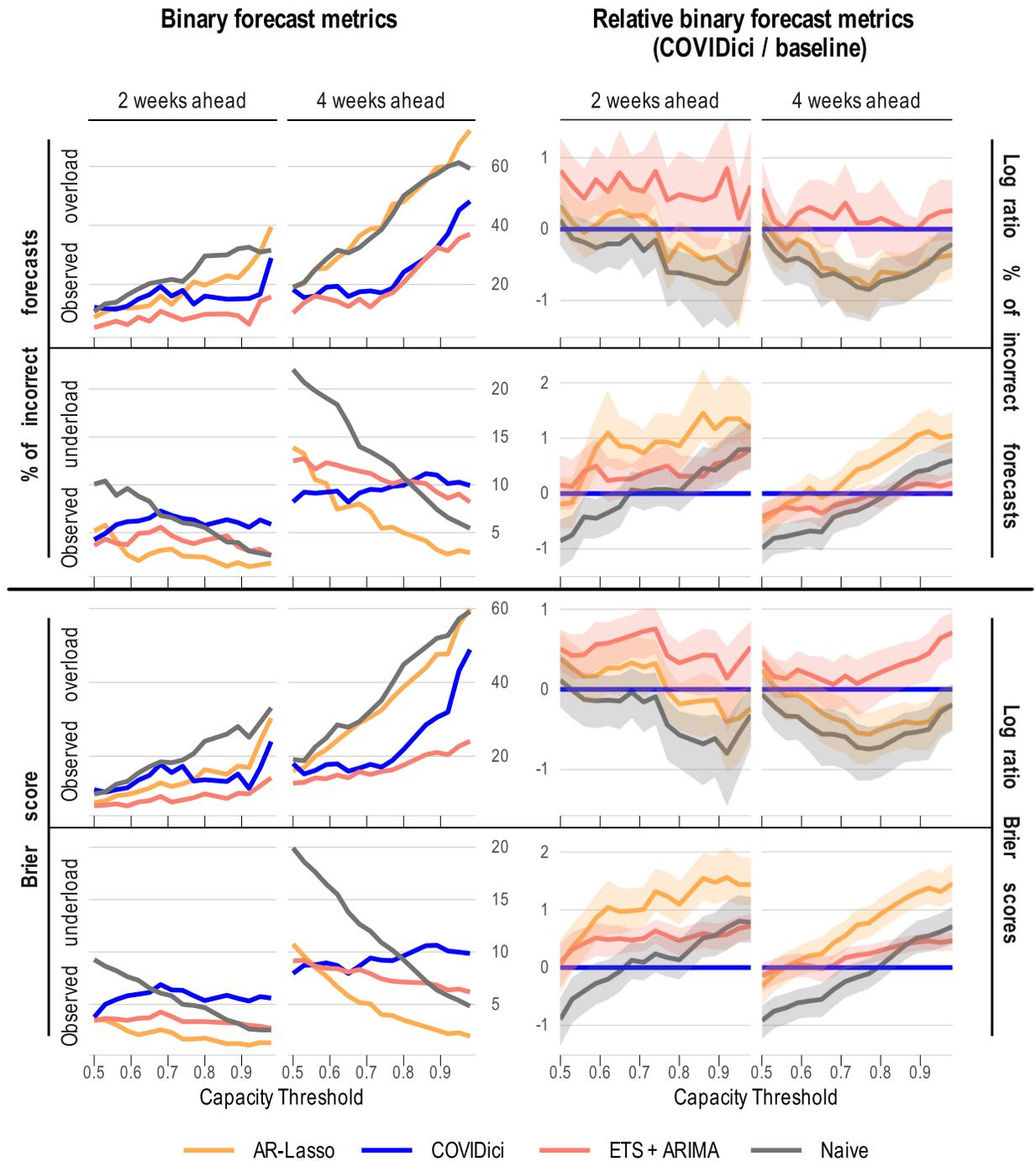
Mathematical modelling was vital to understanding the COVID-19 epidemic in real-time as it unfolded. In France, it led to the anticipation of hospital dynamics in the first epidemic wave [59] and the emergence of the Delta variant [36]. However, most of these studies were performed at a national level and/or at a single time point, which made their impact limited from a public health point of view. To our knowledge, there were only two continuously running forecasting models in France in 2021 at the subnational level: COVID<sub>i</sub>ci and an ensemble of

Median weighted interval score (WIS) of COVIDici relative to baselines



**Fig 6. Median weighted interval score of COVIDici relative to other models.** Ratio is the median weighted interval score (WIS) for COVIDici divided by median WIS using all other models as baselines. Each 95% confidence interval (CI) is based on the bias-corrected and accelerated (BCa) bootstrap confidence interval generated by nonparametric bootstrap with 10000 replicates. The p value is the smallest alpha such that 1 is not contained in the corresponding 1 - alpha CI. See Fig 4 for region code definitions.

<https://doi.org/10.1371/journal.pcbi.1012124.g006>



**Fig 7. Binary metrics for ICU occupancy overload by capacity threshold.** Metrics are conditioned separately for dates where ICU overload was observed or not observed respectively relative to arbitrary capacity thresholds. Thresholds are defined as a proportion of the peak ICU occupancy observed in a geographic unit. Shaded areas correspond to 95% bias-corrected and accelerated (BCa) confidence intervals based on 10000 replicates of a non-parametric bootstrap that assumes both spatial and temporal independence. Shaded areas completely below the blue line indicate that COVIDici statistically outperformed that model for that metric, threshold and forecast horizon.

<https://doi.org/10.1371/journal.pcbi.1012124.g007>

statistical models implemented by the Pasteur Institute [10]. These two models were built on two types of approaches each with different strengths and limitations. For the latter this meant reasonable forecasting accuracy but only for relatively short forecast horizons.

In addition to real-time sub-national forecasting, a key feature of COVID<sub>i</sub>ci was to offer visualisation of numerous unobservable indicators that can only be inferred through an underlying mechanistic model (e.g. the estimate of all active infections, reproductive number, etc.). This is an asset from a popularisation point of view, but it raises issues because estimates can be strongly biased in areas with low population density.

Furthermore, it is important to distinguish between two types of forecasts. Some, like COVID<sub>i</sub>ci, attempt to project mechanistically what would happen if transmission remains identical, i.e. assuming transmission dynamics equal as those observed in the past weeks. Others, especially the ones built on machine learning, try to extrapolate phenomenologically the pattern of the time series by being sensitive to the most recent changes. In terms of guiding public health decision-making, both are complementary since the former is robust in the long term and can incorporate expected events (such as planned interventions or viral variant replacement), while the latter is accurate on the short term.

## Limitations

It is worthwhile to point out that while this evaluation of COVID<sub>i</sub>ci is retrospective, its development, utilisation and updating occurred entirely during the pandemic. A tremendous number of models were proposed during that time that recommended forecasting various COVID-19 indicators using different approaches, and auxiliary data. However, all such proposals (COVID<sub>i</sub>ci included) lacked the benefit of hindsight and were constrained by implementation feasibility and uncertainty over how well they would generalise into future stages of the pandemic. Thus, we emphasise that the following four limitations identified by our evaluation should be interpreted from this practical perspective.

1. The prediction intervals were far too narrow which obfuscated any performative advantages that it had when considering distributional evaluation metrics (e.g. WIS and Brier score). Future attention must be given to creating a better calibrated prediction interval under acceptable computational expense.
2. Fitting COVID<sub>i</sub>ci both nationally and sub-nationally, for 13 regions and 101 departments, using MCMC is very computationally expensive. At the time the COVID<sub>i</sub>ci was ended, the model fitting update on a high-performance cluster using 115 cores took over 4 hours to complete. This meant that refitting the model was rarely feasible from a practical perspective, which is problematic when trying to incorporate additional model fits for different scenarios or it is discovered that a server had been down for maintenance or that a programming oversight led to only saving 3 forecast quantiles.
3. COVID<sub>i</sub>ci exhibited relatively weak performance up to the two-week horizon to predict ICU occupancy compared to statistical modelling approaches. However, this was expected as COVID<sub>i</sub>ci only uses hospital data to update its inference and there is a nearly two-week delay between infection and hospital admission [60]. This decreased performance at the one- and two-week horizons relative to a benchmark is also consistent with findings regarding other mechanistic compartmental models predicting cumulative deaths at the national level in the United States [61].
4. From a pure forecasting perspective, COVID<sub>i</sub>ci only has modest improvements at the four-week horizon compared to rather basic baseline models such as the Naive model. We

emphasise though that this is not the whole story because, contrary to statistical models, it provides a full epidemiological insight in the form of unobservable estimated parameters such as spatialised current prevalence and attack rates.

Regarding limitations to our retrospective evaluation, the results found here may not be immediately generalizable everywhere. The implementation of mechanistic compartmental models requires many simplifying biological assumptions that depend on expert opinion which exists in the context of what is known about the disease at that time. If COVIDici were to be built again today one would likely not use all the same assumptions, e.g. full immunity after recovery from infection. Furthermore, the effects of viral evolution, governmental interventions and spontaneous social changes are very dependent on the time period during which the evaluation is framed. As a result it is hard to comment on the reproducibility of these results for other locations at different times.

## Conclusion

Many countries are decreasing their investment in epidemic surveillance, and some rely on statistical model forecasting with the inclusion of new predictors, such as that from wastewater data. However, there is still room for compartmental model forecasts like COVIDici that can rely on variables with a high level of sampling such as hospital admissions data.

Regarding SARS-CoV-2, future extensions of COVIDici would require updating the model to account more precisely for the diversity in immune protection among individuals, given the number of natural infections since the evolution of the Omicron variants [17]. However, existing non-Markovian models suggest that this is feasible [60]. In several ways COVIDici's forecasting potential was under-exploited. It already incorporated variations in vaccine coverage but, thanks to its mechanistic nature, it could also readily include planned events such as school holidays or early predictors of variations in reproduction numbers such as viral evolution or weather.

The main strength of COVIDici was improved accuracy at the four-week horizon for point forecasts of ICU occupancy compared to the statistical baseline models, especially during the trailing edge of waves. For anticipating wave peaks, COVIDici had one of the best overall performances with respect to the trade-off between the correct prediction rate of observed overload and observed underload on the four-week horizon. The baseline model based on machine learning (i.e. AR-Lasso) on the other hand failed to reasonably anticipate overload, despite relatively optimistic performance in terms of more standard metrics for continuous variables such as WIS. This should serve as a cautionary example for similar models that avoid large errors by avoiding large predictions at peaks of the waves. Systematically avoiding pessimistic predictions may improve the resulting evaluation score, but it also may complicate decision-making regarding non-pharmaceutical government interventions which is a common goal when forecasting hospital strain. To detect models exhibiting such undesirable behaviour, it seems reasonable to consider alternative evaluation metrics based on ICU overload (i.e. binarised ICU occupancy) especially when the evaluation period contains frequent interventions in the training sets.

Those building surveillance systems in future pandemics may consider the application of non-Markovian compartmental models despite their limitations. This is evidenced by the fact that it is clearly feasible since COVIDici was successfully developed and deployed in real-time during the COVID-19 pandemic and that long-term forecasting utility, albeit modest, has subsequently been established at the national and sub-national levels for ICU occupancy.

## Supporting information

**S1 Text. Technical details for real-time forecasting of COVID-19 in France using a non-Markovian mechanistic model.**

(PDF)

## Acknowledgments

We want to thank Santé publique France for providing timely open-source data, the South Green computational platform/IRD for access to their cluster and the French Institute of Bioinformatics for support hosting the COVIDici web application.

## Author Contributions

**Conceptualization:** Samuel Alizon, Mircea T. Sofonea.

**Data curation:** Alexander Massey, Corentin Boennec.

**Formal analysis:** Alexander Massey.

**Funding acquisition:** Samuel Alizon, Mircea T. Sofonea.

**Investigation:** Mircea T. Sofonea.

**Methodology:** Mircea T. Sofonea.

**Project administration:** Christophe Blanchet, Samuel Alizon, Mircea T. Sofonea.

**Software:** Alexander Massey, Corentin Boennec, Claudia Ximena Restrepo-Ortiz, Christophe Blanchet, Mircea T. Sofonea.

**Supervision:** Mircea T. Sofonea.

**Validation:** Alexander Massey.

**Visualization:** Alexander Massey, Corentin Boennec.

**Writing – original draft:** Alexander Massey, Corentin Boennec, Samuel Alizon, Mircea T. Sofonea.

**Writing – review & editing:** Alexander Massey, Samuel Alizon, Mircea T. Sofonea.

## References

1. Becker AD, Grantz KH, Hegde ST, Bérubé S, Cummings DAT, Wesolowski A. Development and dissemination of infectious disease dynamic transmission models during the COVID-19 pandemic: what can we learn from other pathogens and how can we move forward? *Lancet Digit Health*. 2021; 3(1): e41–e50. [https://doi.org/10.1016/S2589-7500\(20\)30268-5](https://doi.org/10.1016/S2589-7500(20)30268-5) PMID: 33735068
2. Brooks-Pollock E, Danon L, Jombart T, Pellis L. Modelling that shaped the early COVID-19 pandemic response in the UK. *Philos Trans R Soc Lond B Biol Sci*. 2021; 376(1829):20210001. <https://doi.org/10.1098/rstb.2021.0001> PMID: 34053252
3. Cramer EY, Huang Y, Wang Y, Ray EL, Cornell M, Bracher J, et al. The United States COVID-19 Forecast Hub dataset. *Sci Data*. 2022; 9:462. <https://doi.org/10.1038/s41597-022-01517-w> PMID: 35915104
4. Bracher J, Wolfram D, Deuschel J, Görgen K, Ketterer JL, Ullrich A, et al. A pre-registered short-term forecasting study of COVID-19 in Germany and Poland during the second wave. *Nat Commun*. 2021; 12:5173. <https://doi.org/10.1038/s41467-021-25207-0> PMID: 34453047
5. Sherratt K, Gruson H, Grah R, Johnson H, Niehus R, Prasse B, et al. Predictive performance of multi-model ensemble forecasts of COVID-19 across European nations. *eLife*. 2023; 12:e81916. <https://doi.org/10.7554/eLife.81916> PMID: 37083521

6. Karatayev VA, Anand M, Bauch CT. Local lockdowns outperform global lockdown on the far side of the COVID-19 epidemic curve. *Proc Natl Acad Sci U S A*. 2020; 117:24575–24580. <https://doi.org/10.1073/pnas.2014385117> PMID: 32887803
7. Lynch CJ, Gore R. Short-Range Forecasting of COVID-19 During Early Onset at County, Health District, and State Geographic Levels Using Seven Methods: Comparative Forecasting Study. *J Med Internet Res*. 2021; 23:e24925. <https://doi.org/10.2196/24925> PMID: 33621186
8. Barría-Sandoval C, Ferreira G, Benz-Parra K, López-Flores P. Prediction of confirmed cases of and deaths caused by COVID-19 in Chile through time series techniques: A comparative study. *PLoS One*. 2021; 16:e0245414. <https://doi.org/10.1371/journal.pone.0245414> PMID: 33914758
9. Gecili E, Ziady A, Szczesniak RD. Forecasting COVID-19 confirmed cases, deaths and recoveries: Revisiting established time series modeling through novel applications for the USA and Italy. *PLoS One*. 2021; 16:e0244173. <https://doi.org/10.1371/journal.pone.0244173> PMID: 33411744
10. Paireau J, Andronico A, Hozé N, Layan M, Crépey P, Roumagnac A, et al. An ensemble model based on early predictors to forecast COVID-19 health care demand in France. *Proc Natl Acad Sci U S A*. 2022; 119(18):e2103302119. <https://doi.org/10.1073/pnas.2103302119> PMID: 35476520
11. Keeling MJ, Rohani P. Modeling infectious diseases in humans and animals. Princeton University Press; 2008.
12. Reyné B, Saby N, Sofonea MT. Principles of mathematical epidemiology and compartmental modelling application to COVID-19. *Anaesth Crit Care Pain Med*. 2022; 41(1):101017. <https://doi.org/10.1016/j.accpm.2021.101017> PMID: 34971801
13. Rahimi I, Chen F, Gandomi AH. A review on COVID-19 forecasting models. *Neural Comput Appl*. 2021. <https://doi.org/10.1007/s00521-020-05626-8> PMID: 33564213
14. Gatto A, Accarino G, Aloisi V, Immorlano F, Donato F, Aloisio G. Limits of Compartmental Models and New Opportunities for Machine Learning: A Case Study to Forecast the Second Wave of COVID-19 Hospitalizations in Lombardy, Italy. *Informatics (MDPI)*. 2021; 8(3).
15. Sofonea MT, Alizon S. Anticipating COVID-19 intensive care unit capacity strain: A look back at epidemiological projections in France. *Anaesth Crit Care Pain Med*. 2021; 40(4):100943. <https://doi.org/10.1016/j.accpm.2021.100943> PMID: 34479681
16. Santé Publique France. Website: Données hospitalières relatives à l'épidémie de COVID-19; 2020. Available from: <https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/>.
17. Sofonea MT, Roquebert B, Foulongne V, Morquin D, Verdurme L, Trombert-Paolantoni S, et al. Analyzing and Modeling the Spread of SARS-CoV-2 Omicron Lineages BA.1 and BA.2, France, September 2021–February 2022. *Emerg Infect Dis*. 2022; 28:1355–1365. <https://doi.org/10.3201/eid2807.220033> PMID: 35642476
18. Casassus B. Elections loom large in France's pandemic policies. *BMJ*. 2022; p. o439. <https://doi.org/10.1136/bmj.o439> PMID: 35273023
19. Sofonea MT, Reyné B, Alizon S. COVIDSIM-FR—Combining statistical analysis of hospital data and parsimonious non Markovian modelling for inferring epidemiological parameters and simulating NPI of the COVID-19 epidemic in France; 2020. Available from: <https://bioinfo-shiny.ird.fr/COVIDSIM2-fr/>.
20. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: <https://www.R-project.org>.
21. Sofonea MT, Reyné B, Elie B, Djidjou-Demasse R, Selinger C, Michalakis Y, et al. Memory is key in capturing COVID-19 epidemiological dynamics. *Epidemics*. 2021; p. 100459. <https://doi.org/10.1016/j.epidem.2021.100459> PMID: 34015676
22. Massey A, Boennec C, Restrepo-Ortiz C, Blanchet C, Alizon S, Sofonea MT. COVIDici: Real-time forecasting of COVID-19-related hospital strain in France using a non-Markovian mechanistic model; 2023. Available from: <https://doi.org/10.5281/zenodo.7641132>.
23. Hansen CH, Michlmayr D, Gubbels SM, Mølbak K, Ethelberg S. Assessment of protection against reinfection with SARS-CoV-2 among 4 million PCR-tested individuals in Denmark in 2020: a population-level observational study. *Lancet*. 2021; 397(10280):1204–1212. [https://doi.org/10.1016/S0140-6736\(21\)00575-4](https://doi.org/10.1016/S0140-6736(21)00575-4) PMID: 33743221
24. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. *Lancet Infect Dis*. 2020; 20(6):669–677. [https://doi.org/10.1016/S1473-3099\(20\)30243-7](https://doi.org/10.1016/S1473-3099(20)30243-7) PMID: 32240634
25. Reyné B, Selinger C, Sofonea MT, Miot S, Pisoni A, Tuaille E, et al. Analysing different exposures identifies that wearing masks and establishing COVID-19 areas reduce secondary-attack risk in aged-care facilities. *Int J Epidemiol*. 2021; 50(6):1788–1794. <https://doi.org/10.1093/ije/dyab121> PMID: 34999872

26. Santé publique France. Données relatives aux personnes vaccinées contre la Covid-19 (VAC-SI)—data.gouv.fr; 2021. Available from: <https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-personnes-vaccinees-contre-la-covid-19-1/>.
27. Haug N, Geyrhofer L, Londei A, Dervic E, Desvars-Larrive A, Loreto V, et al. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat Hum Behav.* 2020; 4(12):1303–1312. <https://doi.org/10.1038/s41562-020-01009-0> PMID: 33199859
28. Haim-Boukobza S, Roquebert B, Trombert-Paolantoni S, Lecorche E, Verdurme L, Foulongne V, et al. Detection of Rapid SARS-CoV-2 Variant Spread, France, January 26–February 16, 2021. *Emerg Infect Dis.* 2021; 27(5):1496–1499. <https://doi.org/10.3201/eid2705.210397> PMID: 33769253
29. Roquebert B, Trombert-Paolantoni S, Haim-Boukobza S, Lecorche E, Verdurme L, Foulongne V, et al. The SARS-CoV-2 B.1.351 lineage (VOC Beta) is outgrowing the B.1.1.7 lineage (VOC Alpha) in some French regions in April 2021. *Euro Surveill.* 2021; 26(23):2100447. <https://doi.org/10.2807/1560-7917.ES.2021.26.23.2100447> PMID: 34114541
30. Alizon S, Haim-Boukobza S, Foulongne V, Verdurme L, Trombert-Paolantoni S, Lecorche E, et al. Rapid spread of the SARS-CoV-2 Delta variant in some French regions, June 2021. *Euro Surveill.* 2021; 26(28):2100573. <https://doi.org/10.2807/1560-7917.ES.2021.26.28.2100573> PMID: 34269174
31. Nishiura H, Linton NM, Akhmetzhanov AR. Serial interval of novel coronavirus (COVID-19) infections. *Int J Infect Dis.* 2020; 93:284–286. <https://doi.org/10.1016/j.ijid.2020.02.060> PMID: 32145466
32. Dagan N, Barda N, Kepten E, Miron O, Perchik S, Katz MA, et al. BNT162b2 mRNA Covid-19 Vaccine in a Nationwide Mass Vaccination Setting. *N Engl J Med.* 2021. <https://doi.org/10.1056/NEJMoa2101765>
33. Santé Publique France. Website: Données relatives aux personnes vaccinées contre la COVID-19; 2020. Available from: <https://www.data.gouv.fr/fr/datasets/donnees-relatives-aux-personnes-vaccinees-contre-la-covid-19-1/>.
34. Santé publique France. Données hospitalières relatives à l'épidémie de COVID-19 (SI-VIC)—data.gouv.fr; 2020. Available from: <https://www.data.gouv.fr/fr/datasets/donnees-hospitalieres-relatives-a-lepidemie-de-covid-19/>.
35. Hartig F, Minunno F, Paul S. BayesianTools: General-Purpose MCMC and SMC Samplers and Tools for Bayesian Statistics; 2020. Available from: <https://CRAN.R-project.org/package=BayesianTools>.
36. Alizon S, Selinger C, Sofonea MT, Haim-Boukobza S, Giannoli JM, Ninove L, et al. Epidemiological and clinical insights from SARS-CoV-2 RT-PCR crossing threshold values, France, January to November 2020. *Euro Surveill.* 2022; 27. <https://doi.org/10.2807/1560-7917.ES.2022.27.6.2100406> PMID: 35144725
37. Lefrant JY, Pirracchio R, Benhamou D, Dureuil B, Pottecher J, Samain E, et al. ICU bed capacity during COVID-19 pandemic in France: From ephemeral beds to continuous and permanent adaptation. *Anaesth Crit Care Pain Med.* 2021; 40:100873. <https://doi.org/10.1016/j.accpm.2021.100873> PMID: 33910085
38. INSEE. Estimation de la population au 1er janvier 2020 | INSEE; 2020. Available from: <https://www.insee.fr/fr/statistiques/1893198>.
39. Hyndman RJ, Athanasopoulos G. Forecasting: principles and practice, 3rd edition. Melbourne, Australia: bookdown; 2021. Available from: <https://OTexts.com/fpp3>.
40. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Series B Stat Methodol.* 1996; 58(1):267–288. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
41. Kuhn M. caret: Classification and Regression Training; 2022. Available from: <https://CRAN.R-project.org/package=caret>.
42. Bracher J, Ray EL, Gneiting T, Reich NG. Evaluating epidemic forecasts in an interval format. *PLoS Comput Biol.* 2021; 17(2):e1008618. <https://doi.org/10.1371/journal.pcbi.1008618> PMID: 33577550
43. Bosse NI, Gruson H, Cori A, van Leeuwen E, Funk S, Abbott S. Evaluating Forecasts with scoringutils in R. *arXiv.* 2022.
44. Lynch CJ, Gore R. Application of one-, three-, and seven-day forecasts during early onset on the COVID-19 epidemic dataset using moving average, autoregressive, autoregressive moving average, autoregressive integrated moving average, and naïve forecasting methods. *Data Brief.* 2021; 35:106759. <https://doi.org/10.1016/j.dib.2021.106759> PMID: 33521186
45. Friedman J, Liu P, Troeger CE, Carter A, Reiner RC, Barber RM, et al. Predictive performance of international COVID-19 mortality forecasting models. *Nat Commun.* 2021; 12:2609. <https://doi.org/10.1038/s41467-021-22457-w> PMID: 33972512
46. Meakin S, Abbott S, Bosse N, Munday J, Gruson H, Hellewell J, et al. Comparative assessment of methods for short-term forecasts of COVID-19 hospital admissions in England at the local level. *BMC Med.* 2022; 20:86. <https://doi.org/10.1186/s12916-022-02271-x> PMID: 35184736

47. Spiliopoulos L. On the effectiveness of COVID-19 restrictions and lockdowns: Pan metron ariston. *BMC Public Health*. 2022; 22(1):1842. <https://doi.org/10.1186/s12889-022-14177-7> PMID: 36183075
48. Haug N, Geyrhofer L, Londei A, Dervic E, Desvars-Larrive A, Loreto V, et al. Ranking the effectiveness of worldwide COVID-19 government interventions. *Nat Hum Behav*. 2020; 4(12):1303–1312. <https://doi.org/10.1038/s41562-020-01009-0> PMID: 33199859
49. Coroneo L, Iacone F, Paccagnini A, Monteiro PS. Testing the predictive accuracy of COVID-19 forecasts. *Int J Forecast*. 2023; 39:606–622. <https://doi.org/10.1016/j.ijforecast.2022.01.005> PMID: 35125573
50. Döhrn R. Comparing forecast accuracy in small samples. RWI—Leibniz-Institut für Wirtschaftsforschung, Ruhr-University Bochum, TU Dortmund University, University of Duisburg-Essen; 2019. 833. Available from: <https://ideas.repec.org/p/zbw/rwirep/833.html>.
51. Huang Y, Xu H, Calian V, Hsu JC. To permute or not to permute. *Bioinformatics*. 2006; 22:2244–2248. <https://doi.org/10.1093/bioinformatics/btl383> PMID: 16870938
52. Desgagné A, Castilloux AM, Angers JF, Lorier JL. The Use of the Bootstrap Statistical Method for the Pharmacoeconomic Cost Analysis of Skewed Data. *Pharmacoeconomics*. 1998; 13:487–497. <https://doi.org/10.2165/00019053-199813050-00002> PMID: 10180748
53. Tukey JW. *Exploratory Data Analysis*. Addison-Wesley; 1977.
54. Rubin M, Donkin C. Exploratory hypothesis tests can be more compelling than confirmatory hypothesis tests. *Philos Psychol*. 2022; p. 1–29. <https://doi.org/10.1080/09515089.2022.2113771>
55. Efron B. Better Bootstrap Confidence Intervals. *J Am Stat Assoc*. 1987; 82:171–185. <https://doi.org/10.1080/01621459.1987.10478410>
56. Helwig NE. nptest: Nonparametric Bootstrap and Permutation Tests; 2023. Available from: <https://CRAN.R-project.org/package=nptest>.
57. Hall P. *The Bootstrap and Edgeworth Expansion*. Springer; 1992.
58. Thulin M. boot.pval: Bootstrap p-Values; 2023. Available from: <https://CRAN.R-project.org/package=boot.pval>.
59. Salje H, Kiem CT, Lefrancq N, Courtejoie N, Bosetti P, Paireau J, et al. Estimating the burden of SARS-CoV-2 in France; 2020. Available from: <https://hal-pasteur.archives-ouvertes.fr/pasteur-02548181>.
60. Reyné B, Richard Q, Selinger C, Sofonea MT, Djidjou-Demasse R, Alizon S. Non-Markovian modelling highlights the importance of age structure on Covid-19 epidemiological dynamics. *Math Model Nat Phenom*. 2022; 17:7. <https://doi.org/10.1051/mmnp/2022008>
61. Coroneo L, Iacone F, Paccagnini A, Santos Monteiro P. Testing the predictive accuracy of COVID-19 forecasts. *Int J Forecast*. 2023; 39(2):606–622. <https://doi.org/10.1016/j.ijforecast.2022.01.005> PMID: 35125573