




Research and Applications

The benefit of augmenting open data with clinical data-warehouse EHR for forecasting SARS-CoV-2 hospitalizations in Bordeaux area, France

Thomas Ferté ^{1,2,3}, Vianney Jouhet ^{1,3}, Romain Griffier^{1,3}, Boris P. Hejblum ^{2,3}, and Rodolphe Thiébaud^{1,2,3}; on behalf of the Bordeaux University Hospital Covid-19 Crisis Task Force

¹Service d'Information Médicale, CHU Bordeaux, Bordeaux, France, ²INRIA Bordeaux Sud Ouest, équipe SISTM, Talence, France and ³Centre Inserm Bordeaux Population Health, University of Bordeaux, UMR 1219, Bordeaux, France

Preprint available at: Ferté, Thomas and Jouhet, Vianney and Griffier, Romain and Hejblum, Boris P. and Thiébaud, Rodolphe and crisis Task, Bordeaux University Hospital Covid-19, The Benefit of Augmenting Open Data with Clinical Data-Warehouse EHR for Forecasting SARS-CoV-2 Hospitalizations in Bordeaux Area, France. Available at SSRN: <https://ssrn.com/abstract=4071506> or <http://dx.doi.org/10.2139/ssrn.4071506>.

Corresponding Author: Rodolphe Thiébaud, MD, PhD, ISPED, Université de Bordeaux, 146 Rue Léo Saignat, 33000 Bordeaux, France; rodolphe.thiebaud@u-bordeaux.fr

Received 27 May 2022; Revised 12 July 2022; Editorial Decision 19 September 2022; Accepted 19 October 2022

ABSTRACT

Objective: The aim of this study was to develop an accurate regional forecast algorithm to predict the number of hospitalized patients and to assess the benefit of the Electronic Health Records (EHR) information to perform those predictions.

Materials and Methods: Aggregated data from SARS-CoV-2 and weather public database and data warehouse of the Bordeaux hospital were extracted from May 16, 2020 to January 17, 2022. The outcomes were the number of hospitalized patients in the Bordeaux Hospital at 7 and 14 days. We compared the performance of different data sources, feature engineering, and machine learning models.

Results: During the period of 88 weeks, 2561 hospitalizations due to COVID-19 were recorded at the Bordeaux Hospital. The model achieving the best performance was an elastic-net penalized linear regression using all available data with a median relative error at 7 and 14 days of 0.136 [0.063; 0.223] and 0.198 [0.105; 0.302] hospitalizations, respectively. Electronic health records (EHRs) from the hospital data warehouse improved median relative error at 7 and 14 days by 10.9% and 19.8%, respectively. Graphical evaluation showed remaining forecast error was mainly due to delay in slope shift detection.

Discussion: Forecast model showed overall good performance both at 7 and 14 days which were improved by the addition of the data from Bordeaux Hospital data warehouse.

Conclusions: The development of hospital data warehouse might help to get more specific and faster information than traditional surveillance system, which in turn will help to improve epidemic forecasting at a larger and finer scale.

Key words: SARS-CoV-2, forecasting, electronic health records, data warehouse, machine learning

LAY SUMMARY

The objective of this work was to develop a forecast algorithm to predict the number of hospitalized patients at Bordeaux Hospital. In addition, we assessed the benefit of the Electronic Health Records (EHRs) information to perform those predictions. To perform this task, we used data between May 16, 2020, and January 17, 2022, from national database on SARS-CoV-2 epidemics, public database on weather and the data warehouse of the Bordeaux hospital. The outcomes were the number of hospitalized patients in the Bordeaux Hospital at 7 and 14 days. During the period of 88 weeks, 2561 hospitalizations due to COVID-19 were recorded at the Bordeaux Hospital. The best model had an error of 13.6% at 7 days and 19.8% at 14 days. EHRs from the hospital data warehouse improved the performance by 10% at 7 days and 20% at 14 days. Graphical evaluation showed remaining forecast error was mainly due to delay in slope shift detection. Forecast model showed overall good performance which were improved by the addition of EHRs data. The development of hospital data warehouse might help to get more specific and faster information than traditional surveillance system, which in turn will help to improve epidemic forecasting at a larger and finer scale.

BACKGROUND AND SIGNIFICANCE

Since the end of 2020, millions of SARS-CoV-2 cases have been reported worldwide.^{1–3} This pandemic has had a major impact on health care with an increase of the hospitalizations leading to modifications of the organization of care and to unprecedented population lockdowns to avoid health care system saturation.^{4–6} The ability to anticipate the evolution of the epidemic at a local level is critical to manage the health care system.

To achieve this goal, several forecasting algorithms have been proposed.^{7,8} None was fully satisfactory. Cramer et al⁸ compared different approaches, including regression, compartmental, ensemble, deep-learning, to forecast the number of death related to COVID-19 in the United States. Best models included ensemble, deep learning, and several compartmental methods. They all used epidemiological data and, depending on the model, mobility, and demographics data. Although the performances varied from algorithm to the other, the data used differed between best models without clear trend indicating which data would provide the best performance.

Because many factors change over time (eg, population behavior, government policies, vaccine coverage, virus strain), long-term forecast of COVID-19 is impossible. In France, several approaches to short-term epidemic forecast have been proposed relying on linear algebra,⁹ ensemble methods¹⁰ or neural networks.^{11,12} All models aimed to forecast hospitalizations (among other things), except the one proposed by Carvalho et al,¹¹ which focused on cases, ICU, and deaths. The data used were hospitalizations, Reverse Transcriptase Polymerase Chain Reaction (RT-PCR), Intensive Care Units (ICU), weather, mobility, vaccination, variants of concern, and mask wearing policy data. The approach proposed by Mohimont et al, based on Convolutional Neural Networks seems to achieve the best results with a normalized root mean square deviation at 14 days of 3.2% compared to a mean absolute percentage error of 20% of the approach proposed by Paireau et al at the national level. However, the periods of evaluation are different (May 2021 for Mohimont et al,¹² March to July 2021 for Paireau et al¹⁰) and direct comparison is difficult. In addition, graphical evaluation showed struggle to anticipate slope shift, that is a change in the dynamic of the epidemic toward an increase or a decrease of the number of cases. In addition, the performance improvement added by the different data sources was not formally evaluated. Finally, no model was satisfactory and valid enough to implement immediately.

Previous work focused mainly on national^{9–12} or regional^{10,12} forecast. Yet, finer granularities are needed to inform on local epidemic evolutions. Because hospitals are key actors during this pan-

demic and their saturation is a critical factor of sanitary policies, they are a relevant scale for local forecasting. Since November 2017, the Bordeaux University Hospital has developed a data warehouse based on i2b2 architecture.¹³ It facilitates the use of electronic health records and allows extracting detailed information of the epidemic such as the emergency units and ambulance service notes. We hypothesized that those local data should improve the forecast of the SARS-CoV-2 epidemic.

OBJECTIVE

The objective of this work was to develop an accurate regional forecast algorithm to predict the number of hospitalized patients and to assess the benefit of the Electronic Health Records (EHR) information to perform those predictions.

MATERIALS AND METHODS

Aggregated data from May 16, 2020, to January 17, 2022, regarding French COVID-19 epidemic were included. In order to improve forecasting, several data sources were used.

Open data

Open data included both epidemiologic data from Santé Publique France and weather data from National Oceanic and Atmospheric Administration (NOAA) Integrated Surface Database.^{14,15} Both provide department aggregated data and are daily updated.

Santé Publique France data included hospitalizations, number of RT-PCR, positive RT-PCR, proportion of positive RT-PCR, dominant variant, and number of first dose vaccinated. RT-PCR data were available by age and were grouped as 0–19, 20–59, and 60 and more years old categories. Variant identification data before February 18, 2021 were not available, and majority variant before that date was assigned to wild type.

NOAA data, including temperature, wind speed, humidity, and dew point, were extracted and the Predict Index for COVID-19 Climate Transmissibility—*Index PREDICT de Transmissivité Climatique de la COVID-19*—(IPTCC) was computed.¹⁵ Missing weather data were imputed using a 2-step procedure: (1) the mean value of the adjacent department was imputed; (2) remaining missing values were imputed using last observation carried forward.

EHR data

The Bordeaux Hospital is a large structure comprising 3 hospital structures taking care of nearly 250 000 hospitalized patients and

100 000 emergency consultations during 2020.¹⁶ A data warehouse based on i2b2 structure was built in 2017.¹³ This star architecture is based on a central fact table where each row represents a diagnosis, a laboratory result, a procedure, a medical observation, etc. Each fact is related to other tables with information about the patient, the visit, the provider, or type of fact.¹⁷ This structure allows for quick data queries compared to the usual siloed organizations.

To perform those queries, ontology alignment have been performed on laboratory results, and ad hoc natural language processing tools have been developed, including ROMEDI (ie, a French drug terminology to extract drug information from text), IAMsystem (ie, a dictionary-based approach for name entity recognition), and SmartCRF (ie, a software to visualize and annotate EHR).^{18–20} Several applied projects were performed using those tools including automatic detection of surgical site infection and transfusion associated circulatory overload.^{21,22}

Thanks to those previous experiences, the Bordeaux hospital data warehouse was used, during the pandemic, to describe the current state of the epidemic at the hospital level on a daily basis. Those data were then used in the forecast model including: hospitalizations, hospital and ICU admission and discharge, ambulance service notes, and emergency unit notes. Concepts related to COVID-19 were extracted from notes by dictionary-based approaches (eg, cough, dyspnea, COVID-19). Dictionaries were manually created based on manual chart review to identify terms used by practitioners. Then, the number and proportion of ambulance service calls or hospitalization in emergency units mentioning concepts related to COVID-19 were extracted. Detail of features is available in [Supplementary Table S1](#).

Due to different data acquisition mechanisms, there was a delay between the occurrence of events and the data acquisition. It was of 1 day for EHR data, 5 days for department hospitalizations and RT-PCR, 4 days for weather, 2 days for variants, and 4 days for vaccination. For the training and evaluation of the model, the chosen date was the date of data availability to mimic a real-time streaming forecast.

Statistical and machine learning models

The outcomes were the number of hospitalized patients with SARS-CoV-2 infection in the Bordeaux hospital at 7 and 14 days. Several statistical models and machine learning algorithms were compared: linear regression and Poisson regression with elastic-net penalization, random forest, and Fréchet random forest (ie, a random forest derived method able to learn directly from time series).²³ Negative predicted values were forced to 0.

Modeling strategy

To train the model, the primary analysis used a matrix where each row corresponds to a day and each column to a feature from Gironde (ie, the Bordeaux department) open source and Bordeaux Hospital EHR data. Other department data were not added to limit the number of features. Prediction performance were evaluated depending on the data used for the forecast, initially using only hospitalization and RT-PCR and progressively adding SARS-CoV-2 incidence in Gironde, weather, EHRs, Vaccine, and Variant data. Because of the elastic-net penalization,²⁴ each day the model might select different features by shrinking beta coefficient of unimportant features to zero.

An additional analysis considered the Bordeaux Hospital as an additional unit among the other French departments. The advantage

was to leverage the information from all the departments; the drawback was the impossibility to include specific information only available for the Bordeaux Hospital. Of note, the incorporation as a third level inside the Gironde department would be possible in theory but it would lead to the same restrictions and it would not be feasible in absence of other EHR data from other departments. Results from this additional analysis are available in [Supplementary Table S2](#).

Several feature engineering transformations were performed. The mean, minimum, and maximum value over the last 7 days were computed for each feature, as well as the first derivatives over the last 3, 7, 10, and 14 days. Features were smoothed using a local polynomial regression with a span of either 0 (ie, no smoothing), 7, 14, or 21 days to take into account outliers and weekly variations.

Model evaluation

The models were evaluated every day on the data available from December 1, 2020 to January 17, 2022 (ie, data from May 16, 2020 to December 1, 2020 were used for training only). The models were trained using all prior data available at a given date d and the forecast of the number of hospitalizations at both 7 and 14 days after d was evaluated. Prediction performances were evaluated according to median absolute error (MAE) and median relative error (MRE). Median was chosen over mean because it is less sensitive to extreme values which are frequent when the observed outcome is low (ie, a small absolute error could result in a huge relative error). Bootstrapped 95% confidence interval was provided with 500 samples (more details in the [Supplementary Material](#)). Graphical evaluation was also performed. Prediction intervals were estimated by bootstrap and also with an operational rule of more or less 10% and 20% of the predicted value at 7 and 14 days. For the latter, given $\widehat{\text{hosp}}$ the predicted hospitalization, the prediction interval at 7 and 14 days were $[0.9 \times \widehat{\text{hosp}} ; 1.1 \times \widehat{\text{hosp}}]$ and $[0.8 \times \widehat{\text{hosp}} ; 1.2 \times \widehat{\text{hosp}}]$, respectively.

RESULTS

Description

Relationship between predictors and hospitalizations change over time. For instance, as described in [Supplementary Figure S1](#), both RT-PCR in the Bordeaux Hospital and in Gironde well anticipated hospitalizations from June 2020 to December 2020. Unfortunately, the relationship become less consistent beyond December 2020. For instance, Gironde RT-PCR are synchronous to hospitalization peak in April 2021 but do not anticipate it. In addition, the large increase of positive RT-PCR during the end of the summer 2021 is not associated with a similar increase of hospitalizations, which is probably due to vaccination. Those findings explain why the forecast of SARS-CoV-2 hospitalizations is difficult and why it is interesting to leverage different data sources to improve prediction capacity.

Model performance

[Table 1](#) shows the forecast performance depending on the features included in the model. A simple model including the hospitalizations and the RT-PCR led to a MAE of 9.04 and MRE of 17.6%, that is a prediction up to 17.6% more or less than the observed number of hospitalization 7 days later. The addition of RT-PCR and hospitalizations from the Gironde area slightly decreased MAE at 7 days by 0.25, increased it at 14 days by 1.33 and decreased MRE by respectively 0.5% and 1.6% at 7 and 14 days. Adding weather data

Table 1. Forecast performance by data source

Forecast	Data [number of features]	MAE [95% CI]	MRE [95% CI]
7 days	Hosp + RT-PCR [266]	9.04 [6.36; 11.88]	0.176 [0.116; 0.243]
7 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR [526]	8.79 [6.41; 11.19]	0.171 [0.119; 0.227]
7 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather [666]	7.51 [4.72; 10.38]	0.156 [0.097; 0.222]
7 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service [2986]	6.59 [3.05; 9.78]	0.139 [0.065; 0.222]
7 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service + variants + vaccine [2990]	6.67 [3.08; 9.75]	0.136 [0.063; 0.223]
7 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service + variants + vaccine + interaction (vaccine × RT-PCR) [3470]	6.12 [2.12; 9.89]	0.125 [0.048; 0.218]
14 days	Hosp + RT-PCR [266]	15.88 [9.30; 21.56]	0.334 [0.204; 0.435]
14 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR [526]	17.21 [9.92; 24.99]	0.318 [0.187; 0.458]
14 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather [666]	11.63 [6.90; 17.16]	0.242 [0.143; 0.339]
14 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service [2986]	9.48 [4.92; 15.46]	0.194 [0.103; 0.304]
14 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service + variants + vaccine [2990]	9.37 [4.85; 15.34]	0.198 [0.105; 0.302]
14 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service + variants + vaccine + interaction (vaccine × RT-PCR) [3470]	9.74 [4.67; 16.87]	0.212 [0.097; 0.329]
11 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service + variants + vaccine [2990]	8.75 [4.13; 13.37]	0.185 [0.091; 0.298]
18 days	Hosp + RT-PCR + Gironde hosp + Gironde RT-PCR + weather + CT-scan + emergency units + ambulance service + variants + vaccine [2990]	11.97 [5.51; 18.45]	0.235 [0.115; 0.350]

Note: In bold, the best model per forecast time frame.

MAE: median absolute error; MRE: median relative error.

improved 7- and 14-day MAE by respectively 1.28 and 5.58 hospitalizations. The addition of specific EHR data from the hospital data warehouse improved the forecast at both 7 and 14 days by respectively 0.92 and 2.15 in term of MAE, and by 1.7% and 4.8% in term of MRE. As both emergency units and ambulance service data provide information about the symptomatic patients who are different from the positive patients detected by RT-PCR and the severe patients detected by hospitalization, they might help to better anticipate the evolution of the epidemic. Figure 2C supports those findings and shows that using EHR data anticipates the April peak, better forecasts the 2021–2022 winter increase and is more robust to the hospital cluster of December 2020. Last, we evaluated the information added by the number of vaccinated people and the variants distribution. Those additional data provided similar forecast performance and vaccine-RT-PCR interaction improved forecast at 7 days but decreased it at 14 days. Similar results were obtained with the additional analysis considering the Bordeaux Hospital as an additional department; results are available at Supplementary Table S2. Overall, we considered the model including all the data but without the vaccine-RT-PCR interaction as the best model.

Information retrieval from data warehouse was done with a 1-day delay whereas it was a 5-day delay for hospitalizations and RT-PCR at the department level. Therefore, forecasting using data from Bordeaux Hospital data warehouse gives a 4-day advantage over usual open data. To evaluate its consequences, we compared 7-day forecast with 11-day forecast and 14-day forecast with 18-day forecast. As shown in Table 1, this resulted in a MAE and MRE decrease of respectively 2.08% and 4.9% at 7 days and of respectively 2.6% and 3.7% at 14 days.

Table 2 shows the performance according to feature engineering and statistical model. Model performance were improved by smoothing and by feature transformation (ie, mean, minimum, max-

imum, and first derivative) of respectively 1.67 and 0.83 absolute error at 7 days and 2.47 and 3.24 absolute error at 14 days. Linear regression with elastic-net penalization outperformed random forest, Fréchet random forest, and Poisson regression with elastic-net penalization. Finally, the best model was the elastic-net model using hospitalizations, RT-PCR, weather, vaccine, variant, emergency units, ambulance service data, vaccine, and majority variant with a smoothing span of 21 days and the mean, min, max, and first derivative feature transformation.

Figure 1 represents the feature importance of the model in term of percentage of selection over the days with asymptotic confidence intervals. At 7 days, the most important features were the ones related to hospitalization, as there is an intrinsic dynamic of hospitalizations, which is less susceptible to be influenced by external factors on such a short period. At 14 days, the most important features were related to RT-PCR in the 60+ years old group, weather data, emergency units, and hospitalizations.

Figure 2A shows the best model predictions from 1 to 14 days. Predictions were mostly accurate except: (1) in December 2020 during the hospital nosocomial cluster at Bordeaux Hospital which had a specific dynamic, (2) in the end of March 2021 where the April decrease is anticipated 2 weeks earlier, and (3) during the summer 2021 and the winter 2021–2022 where the forecast is overestimating hospitalizations partly because the RT-PCR increased massively whereas the hospitalizations increased moderately as depicted in Supplementary Figure S1. The latter might be a consequence of the vaccination campaign and the omicron spread. Figure 2B shows the prediction intervals of the forecast using an ad hoc rule of respectively 20% and 40% prediction interval at 7 and 14 days, which has better coverage percentage than bootstrapped prediction intervals available in the Supplementary Material. This figure shows that prediction intervals are mostly correct except during summer 2021 and

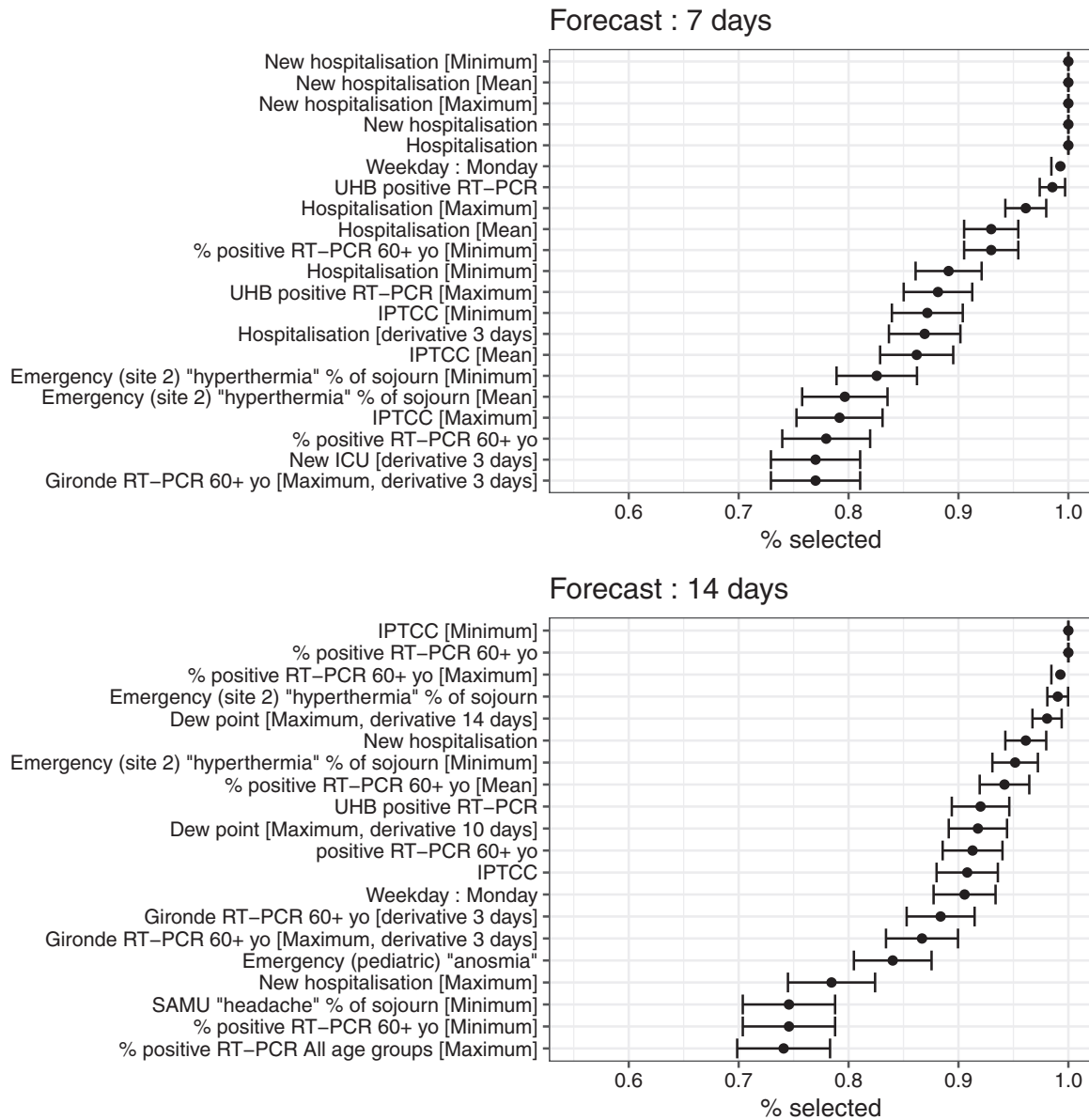


Figure 1. Top 20 features importance with 95% confidence interval of the best model. As model is retrained every day and is elastic-net penalized, each dot represents the proportion of days for which the feature is selected. SAMU—French ambulance service.

winter 2021–2022. The former is explained by the low number of infected—which generates narrow prediction intervals—and the introduction of vaccination which biases the prediction forecast. The latter is explained by the overestimation of hospitalization rise.

Model evolution over time

Figure 3 shows the evolution of MAE and MRE at 7 and 14 days over time. Globally, both MAE and MRE tend to be higher in December 2020, which corresponds to an intrahospital cluster. They also increase after June 2021, which is probably related to the impact of the vaccination campaign followed by the rise of omicron variant.

Selected features by elastic-net penalization change over time. At 14-day forecast, we observe that features related to UHB positive PCR (especially in the elderly), hospitalizations, and IPTCC are frequently selected during all the period. Of note, features related to

pediatric emergency are more often selected after August 2021. At 7-day forecast, the most frequently selected features are less variable over time, with the RT-PCR in the elderly, the hospitalizations, and the IPTCC frequently selected. More details are available in Supplementary Figures S5 and S6.

DISCUSSION

This work demonstrates good overall forecast ability, both in term of relative and absolute error of predictions which were both improved by the addition of weather and data-warehouse information. In addition, the intrinsic “rules” governing the epidemic evolve over time, and external interventions such as vaccination make the hospitalization time series nonstationary, which is an important challenge for a data driven approach. To mitigate this, we used an adaptive approach where the model is trained every day on histori-

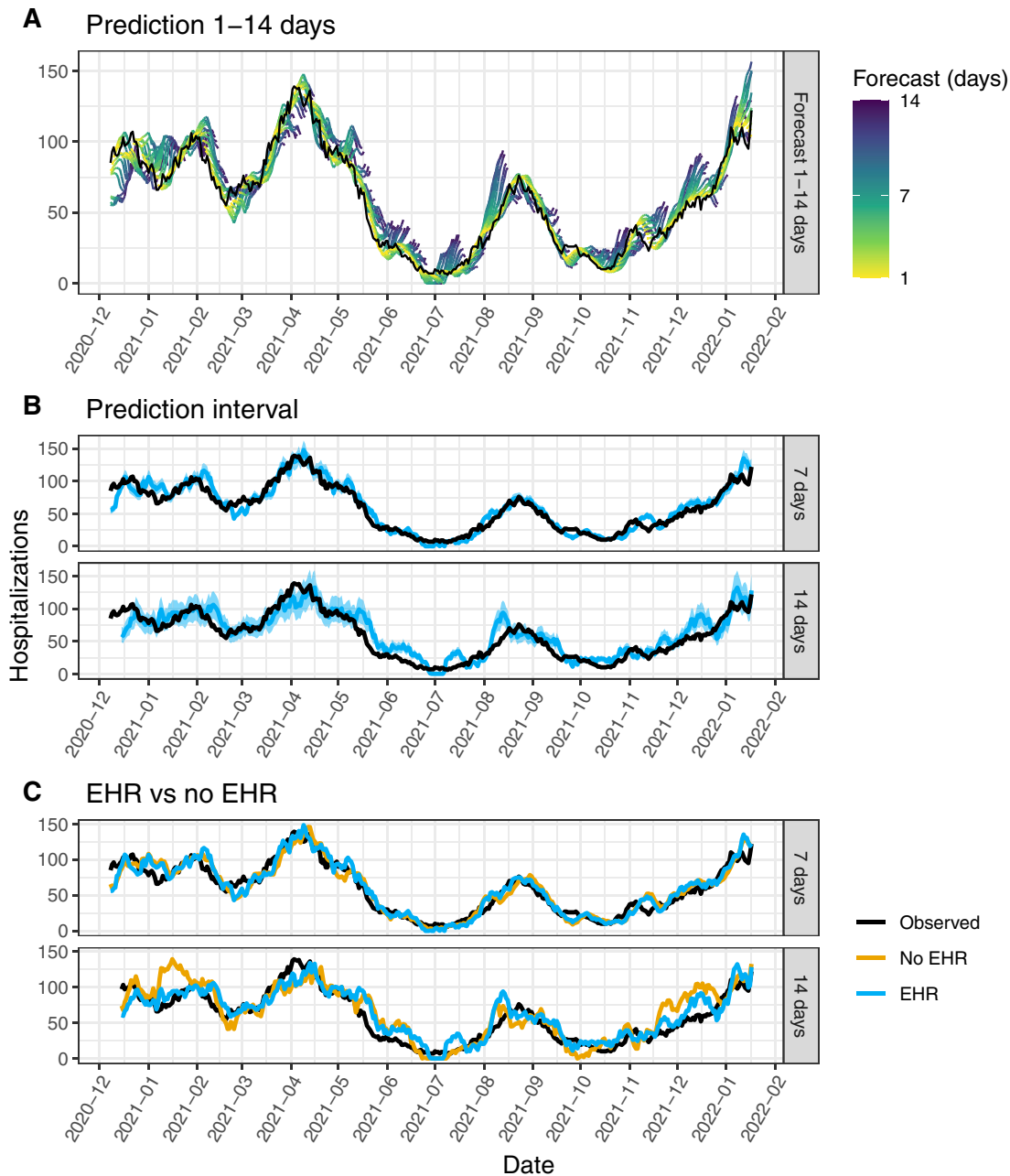


Figure 2. Forecast of best model at 7 and 14 days. (A) The best model forecast from 1 to 14 days. Each string corresponds to the forecast from 1 (yellow) to 14 days (purple) at each day. (B) The 20% prediction intervals at 7 days and 40% at 14 days. Image (C) compares the forecast without EHR data (ie, hospitalizations+RT-PCR+Gironde hospitalizations+Gironde RT-PCR+weather+variant+vaccine).

cal data, which is realistic given the swiftness of penalized linear regression training. In addition, elastic-net penalization allows the model to select different features each day enhancing the adaptiveness of the model.

In their review, Rahim et al⁷ identify compartmental models and deep learning as the most common approaches to perform short-term forecast, followed by machine learning and statistical learning. Formal comparisons of their performance is difficult because dates and locations vary across the reviewed articles. Table 3 summarize findings of selected related work. Cramer et al⁸ describe the COVID-19 Forecast Hub, an open platform where several academic, industry, and independent groups proposed forecasting

model for US SARS-CoV-2 cases, hospitalizations, and deaths, which permits head-to-head comparison. In their paper, they evaluated death forecast performance. They show that performances varied from one algorithm to another, showing no trend as to which data would consistently provide the best performance. Interestingly, the best algorithm used ensemble method, second and third best algorithms used compartmental method.

In France, forecast performance of previous studies are generally claimed to be good as depicted at Table 3.^{9,10,12} However, benchmarking and comparisons are difficult due to different time period and geographical scales. Graphical evaluations help to qualitatively compare model behavior, especially considering slope shifts, but it is

Table 2. Forecast performance depending on modeling hyperparameters

Forecast	Feature transformation	Smoothing span	Machine-learning model	MAE [95% CI]	MRE [95% CI]
7 days	Raw, Mean, Min Max, Derivative	0 days	Linear elastic-net	8.34 [3.13; 13.21]	0.168 [0.063; 0.284]
7 days	Raw, Mean, Min Max, Derivative	7 days	Linear elastic-net	8.23 [2.72; 13.24]	0.169 [0.057; 0.284]
7 days	Raw, Mean, Min Max, Derivative	14 days	Linear elastic-net	7.00 [2.42; 11.34]	0.162 [0.057; 0.263]
7 days	Raw, Mean, Min Max, Derivative	21 days	Linear elastic-net	6.67 [3.08; 9.75]	0.136 [0.063; 0.223]
7 days	Raw	21 days	Linear elastic-net	7.50 [0.90; 17.02]	0.151 [0.021; 0.331]
7 days	Raw, Mean, Min Max, Derivative	21 days	Poisson elastic-net	11.50 [5.75; 19.37]	0.262 [0.144; 0.419]
7 days	Raw, Mean, Min Max, Derivative	21 days	Random Forest	8.00 [4.00; 13.00]	0.173 [0.088; 0.265]
7 days	14-Day raw curves	21 days	Fréchet Forest	7.62 [3.44; 12.11]	0.169 [0.084; 0.274]
14 days	Raw, Mean, Min Max, Derivative	0 days	Linear elastic-net	11.84 [5.62; 18.43]	0.249 [0.107; 0.363]
14 days	Raw, Mean, Min Max, Derivative	7 days	Linear elastic-net	12.79 [6.78; 19.25]	0.267 [0.123; 0.387]
14 days	Raw, Mean, Min Max, Derivative	14 days	Linear elastic-net	10.75 [5.29; 16.78]	0.217 [0.097; 0.344]
14 days	Raw, Mean, Min Max, Derivative	21 days	Linear elastic-net	9.37 [4.85; 15.34]	0.198 [0.105; 0.302]
14 days	Raw	21 days	Linear elastic-net	12.61 [3.01; 25.47]	0.306 [0.056; 0.580]
14 days	Raw, Mean, Min Max, Derivative	21 days	Poisson elastic-net	15.31 [7.17; 25.73]	0.370 [0.205; 0.552]
14 days	Raw, Mean, Min Max, Derivative	21 days	Random Forest	13.00 [10.00; 18.00]	0.269 [0.190; 0.357]
14 days	14-Day raw curves	21 days	Fréchet Forest	13.52 [7.77; 18.49]	0.232 [0.145; 0.339]

Note: In bold, the best model per forecast time frame.

14-day raw curves: as Fréchet model are able to learn trends directly from curves, they are not provided with transformed features.

Models with “Raw” data (ie, no transformation) are trained on 159 features. Models trained on “Raw, Mean, Min Max, Derivative” are trained on 2990 features.

MAE: median absolute error; MRE: median relative error.

not always provided^{9,12} and can show difficulties to anticipate those.¹⁰ Off note, the work proposed by Mohimont et al¹² was able to anticipate November hospitalization slope shift at the national level.

The main strength of our study was to leverage information from different sources. The addition of weather did improve forecasting which is consistent with previous work.²⁵ The use of specific information from a hospital data warehouse improved the performance both by increasing the amount of information and by updating the information faster than public data sources.

In our work, the vaccine and variants data had little effect on the performance of the model. This was partly expected because vaccine is a monotonous increasing curve and majority variant is relatively stable over time as shown in Supplementary Figure S2. In addition, there might be mismeasurement bias of variant as not every Sars-CoV-2 positive swab is tested for variant identification. Furthermore, the information added by variant and vaccine might already be captured by both the RT-PCR and hospitalization features. We also observed that RT-PCR-variant interaction deteriorated the model performance at 14 days. This was explained by a spurious positive correlation between the interaction term and the hospitalizations. Nevertheless, both vaccine and variant are linked to the SARS-CoV-2 epidemic and the inability to leverage that information in machine learning approach might call for a more mechanistic one.^{25–27}

The moderate impact of the addition of Gironde data to the model might be explained by: (1) the information already being captured by the local data from the Bordeaux Hospital and (2) the data collection being slower at the department level than in the hospital. Data consolidation (ie, the update of prior data thanks to new information) was not taken into account because of lack of availability of data versioning. For instance, hospitalization related to Sars-Cov-2 is sometimes identified in the EHR through billing codes in International Classification of Disease, 10th revision, which are only available at the end of the sojourn. This also might have biased the performance evaluation.

Linear regression with elastic-net penalization outperformed other more complex models such as Fréchet or classical random for-

est. This might be explained by: (1) a linear relationship between the features and (2) the difficulty for random forest to extrapolate features relationship outside of the training set space that a linear regression can handle more naturally.

This work outlined the advantage of adding information from the EHRs to improve the forecast, especially at 14 days and during November and December 2021 as depicted at Figure 2C. This was expected because the EHRs provide additional information about the local evolution of the pandemic through the emergency units and the ambulance service data. Those data inform on the more specific population of COVID-19 symptomatic patients who are different from the overall COVID-19 patients detected by RT-PCR and the severe COVID-19 patients who are hospitalized.

Figure 3 outlined that model performance is not homogeneous. Relative error was generally higher when the number of hospitalizations was low (summer 2021, November 2021) which is expected because the denominator is lower during those periods. Absolute error was higher during December 2020, which is expected because the model did not learn on many observations yet, and there was a hospital cluster at this period. It was also higher during the summer 2021 that might be explained by the beginning of vaccine campaign effect.

Although the model performance can be considered as good, there are some limitations. First, performance tends to deteriorate when there is a sudden change of the hospitalizations dynamic. For instance, in February and March 2021, the decrease and the increase of the hospitalizations were not well anticipated by the model. Second, change of the infectiousness of the virus either due to a mutation or vaccine were not immediately learned by the model. This might explain the overestimate during summer 2021 and it might occur again in the future. Third, the model used for this task is a linear regression and it might not capture complex relations. As discussed before, random forest did not improve forecast but other machine learning methods such as reservoir computing may.²⁸

The model is currently used in the Bordeaux University Hospital on a daily basis to anticipate the evolution of the number of COVID-19 hospitalizations. The forecast is used in conjunction

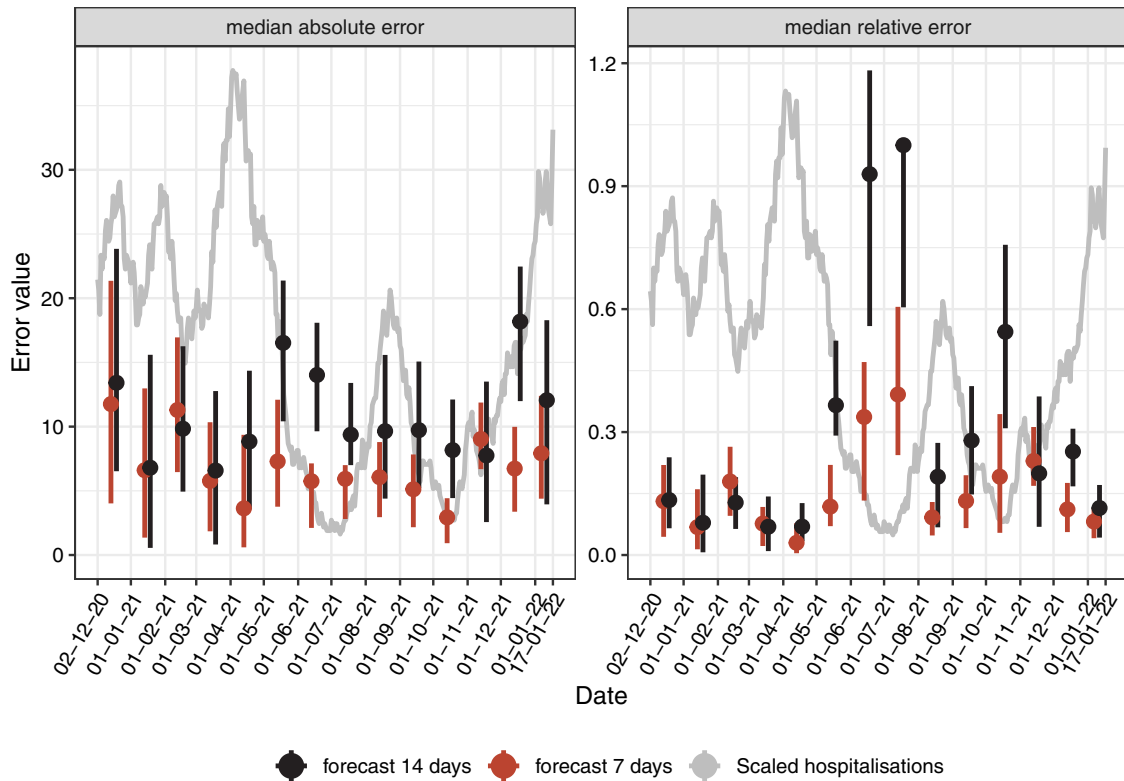


Figure 3. Evolution of MAE and MRE over time at 7 and 14 days. Each dot corresponds to the median absolute error of the corresponding month with the 95% bootstrapped confidence interval.

Table 3. Summary of discussed forecasting methods

Study	Location	Data ^a	Methods	Outcome ^b	Performance
Paireau et al ¹⁰	France	EDMW	Ensemble	H at 14 days	MAPE at 14 days: 20%
Mohimont et al ¹²	France	EDM	Convolutional Neural Networks	H at 14 days	Normalized root mean square deviation at 14 days: 3.2%
Pottier et al ⁹	France	E	Linear algebra	H at 14 days	Average relative error: 12%
Carvalho et al ¹¹	France	E	Neural Networks	C, D, and ICU at 14 days	MAPE: C 4.13%, D 10.26%, ICU 0.92%
Cramer et al ⁸	US	EDM	Ensemble (best)	D at 4 weeks	RMAEvB: 0.66
		E	UMass-MechBayes, Bayesian compartmental model (second best)	D at 4 weeks	RMAEvB: 0.67
		E	Karlen-pypm, discrete time compartmental model (third best)	D at 4 weeks	RMAEvB: 0.70

Note: Relative Mean Absolute Error compared to a baseline model of constant prediction (RMAEvB).

^aE (Epidemiological), D (Demographic), M (Mobility), W (Weather).

^bH (Hospitalizations), D (Death due to COVID-19), ICU (Intensive Care Units), C (Cases).

MAPE: mean absolute percentage error.

with other indicators (number of hospitalizations, RT-PCR, emergency unit's workload, etc.) and is discussed with both clinicians and public health experts. In our experience, the forecast is particularly informative to anticipate when a local peak is reached and the hospitalizations will decrease or the arrival of a new wave.

CONCLUSION

This work highlights the advantage of leveraging several different data sources to improve forecast accuracy. The development of

hospital data warehouse might help to get more specific and faster information than a traditional surveillance system, which in turn will help to improve epidemic forecasting at a larger and finer scale.

FUNDING

This work has been partly supported by Inria, Mission COVID19, GESTEPID project, and Nouvelle Aquitaine regional funding (Prediction territorial COVID N 1333140).

AUTHOR CONTRIBUTIONS

TF contributed to the data curation, formal analysis, software and writing—original draft. VJ contributed to conceptualization, data curation, supervision, funding acquisition, writing—original draft. RG contributed to data curation, software, writing—original draft. BH contributed to conceptualization, methodology, supervision, writing—original draft. RT contributed to conceptualization, methodology, supervision, funding acquisition, writing—original draft.

ETHICAL APPROVAL

No ethics committee approval was needed for this work as data used for modeling were aggregated.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *JAMIA Open* online.

ACKNOWLEDGMENTS

Preprint of this paper is available at SSRN: <https://ssrn.com/abstract=4071506> or <http://dx.doi.org/10.2139/ssrn.4071506>. Code of this paper is available on <https://github.com/thomasferte/PredictCovidOpen.git>. Computer time for this study was provided by the computing facilities of the MCIA (Mésocentre de Calcul Intensif Aquitain). This work has been partly supported by Inria, Mission COVID19, GESTEPID project, and Nouvelle Aquitaine regional funding (Prediction territorial COVID N 1333140). We thank Louis CAPITAINE and Robin GENUER for their help with Fréchet random forest. We thank the Bordeaux University Hospital Covid-19 crisis task force for their help with the data curation and the interpretation: Isabelle FAURE, Philippe REVEL, Eric TENTILLIER, Jean-Michel DINDART, Didier GRUSON, Olivier JOANNES-BOYAU, Jean-Marie Denis MALVY, Thierry PISTONE, Didier NEAU, Duc NGUYEN, Marie-Edith LAFON, Mathieu MOLIMARD, Thierry SCHAEVERBEKE, Nicolas GRENIER, Nathalie SALLES, Francois ROUANET.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY

Open data used for this study are available on data.gouv.fr¹⁴ and ncei.noaa.gov.¹⁵ EHR data used for this study are available upon request. Data available from the Dryad Digital Repository: <https://doi.org/10.5061/dryad.hhmgqnkx>. Code available from github: <https://doi.org/10.5281/zenodo.6595011>.

REFERENCES

- Carrat F, Figoni J, Henny J, *et al.* Evidence of early circulation of SARS-CoV-2 in France: findings from the population-based “CONSTANCES” cohort. *Eur J Epidemiol* 2021; 36 (2): 219–22.
- World Health Organisation. WHO Coronavirus (COVID-19) Dashboard. 2020. <https://covid19.who.int>. Accessed September 27, 2021.
- COVID-19 Cumulative Infection Collaborators. Estimating global, regional, and national daily and cumulative infections with SARS-CoV-2 through Nov 14, 2021: a statistical analysis. *Lancet (London, England)* 2022; 399 (10344): 2351–80.
- Kim G, Wang M, Pan H, *et al.* A health system response to COVID-19 in long-term care and post-acute care: a three-phase approach. *J Am Geriatr Soc* 2020; 68 (6): 1155–61.
- Simões J, Magalhães JPM, Biscaia A, *et al.* Organisation of the State, model of health system and COVID-19 health outcomes in six European countries, during the first months of the COVID-19 epidemic in 2020. *Int J Health Plann Manage* 2021; 36 (5): 1874–86.
- Hübner M, Zingg T, Martin D, *et al.* Surgery for non-Covid-19 patients during the pandemic. *PLoS One* 2020; 15 (10): e0241331.
- Rahimi I, Chen F, Gandomi AH. A review on COVID-19 forecasting models [published online ahead of print February 4, 2021]. *Neural Comput Appl* 2021; 1–11. doi: [10.1007/s00521-020-05626-8](https://doi.org/10.1007/s00521-020-05626-8).
- Cramer EY, Ray EL, Lopez VK, *et al.* Evaluation of individual and ensemble probabilistic forecasts of COVID-19 mortality in the United States. *Proc Natl Acad Sci U S A* 2022; 119 (15): e2113561119.
- Pottier L. Forecast of the Covid19 Epidemic in France. 2021. <https://hal.archives-ouvertes.fr/hal-03183712>. Accessed March 12, 2022.
- Paireau J, Andronico A, Hozé N, *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.
- Carvalho K, Vicente JP, Jakovljevic M, *et al.* Analysis and forecasting incidence, intensive care unit admissions, and projected mortality attributable to COVID-19 in Portugal, the UK, Germany, Italy, and France: predictions for 4 weeks ahead. *Bioengineering* 2021; 8 (6): 84.
- Mohimont L, Chemchem A, Alin F, *et al.* Convolutional neural networks and temporal CNNs for COVID-19 forecasting in France. *Appl Intell* 2021; 51 (12): 8784–809.
- Murphy SN, Mendis ME, Berkowitz DA, *et al.* Integration of clinical and genetic data in the i2b2 architecture. *AMIA Annu Symp Proc AMIA Symp* 2006; 2006: 1040.
- Etalab. Les données relatives au COVID-19 en France – data.gouv.fr. 2020. <https://www.data.gouv.fr/fr/pages/donnees-coronavirus/>. Accessed September 29, 2021.
- Smith A, Lott N, Vose R. The integrated surface database: recent developments and partnerships. *Bull Am Meteorol Soc* 2011; 92 (6): 704–8.
- CHU de Bordeaux. CHU chiffres clés 2020. 2020. <https://www.chu-bordeaux.fr/CHU-de-Bordeaux/Pr%C3%A9sentation-du-CHU/Activit%C3%A9-du-CHU-en-chiffres/CHU-chiffres-clés-2020-planches-.pdf>. Accessed January 3, 2022.
- Post AR, Ai M, Kalsanka Pai A, *et al.* Architecting the data loading process for an i2b2 research data warehouse: full reload versus incremental updating. *AMIA Annu Symp Proc* 2017; 2017: 1411–20.
- Cossin S, Lebrun L, Lobre G, *et al.* Romedi: An Open Data Source About French Drugs on the Semantic Web. *Stud Health Technol Inform* 2019; 264: 79–82.
- Cossin S, Jouhet V, Mougin F, *et al.* IAM at CLEF eHealth 2018: concept annotation and coding in French death certificates [published online ahead of print July 10, 2018]. *ArXiv180703674 Cs*. <http://arxiv.org/abs/1807.03674>. Accessed January 3, 2022.
- Cossin S, Lebrun L, Aymeric N, *et al.* SmartCRF: a prototype to visualize, search and annotate an electronic health record from an i2b2 clinical data warehouse. *Stud Health Technol Inform* 2019; 264: 1445–6.
- Quéroué M, Lashéras-Bauduin A, Jouhet V, *et al.* Automatic detection of surgical site infections from a clinical data warehouse [published online ahead of print September 16, 2019]. *ArXiv190907054 Cs Stat*. <http://arxiv.org/abs/1909.07054>. Accessed January 3, 2022.
- Cossin S, Puntous M, Pujol S, *et al.* Automatic detection of transfusion-associated circulatory overload (TACO) in electronic health records. In: TALMED 2019. Lyon, France; 2019. <https://hal.archives-ouvertes.fr/hal-02288139>. Accessed March 6, 2020.
- Capitaine L, Bigot J, Thiébaud R, *et al.* Fréchet random forests for metric space valued regression with non-euclidean predictors [published online

- ahead of print November 3, 2020]. *ArXiv190601741 Cs Stat.* <http://arxiv.org/abs/1906.01741>. Accessed September 29, 2021.
24. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc B* 2005; 67 (2): 301–20.
 25. Collin A, Hejblum BP, Vignals C, *et al.* Using population based Kalman estimator to model COVID-19 epidemic in France: estimating the effects of non-pharmaceutical interventions on the dynamics of epidemic. *medRxiv* 2021; 2021.07.09.21260259. <https://www.medrxiv.org/content/10.1101/2021.07.09.21260259v3>.
 26. Liu Y, Rocklöv J. The reproductive number of the Delta variant of SARS-CoV-2 is far higher compared to the ancestral SARS-CoV-2 virus. *J Travel Med* 2021; 28 (7): taab124.
 27. Shah ASV, Gribben C, Bishop J, *et al.* Effect of vaccination on transmission of SARS-CoV-2. *N Engl J Med* 2021; 385 (18): 1718–20.
 28. Ghosh S, Senapati A, Mishra A, *et al.* Reservoir computing on epidemic spreading: a case study on COVID-19 cases. *Phys Rev E* 2021; 104 (1–1): 014308.