frontiers | Frontiers in Nephrology

TYPE Original Research
PUBLISHED 02 June 2023
DOI 10.3389/fneph.2023.1179342

OPEN ACCESS

EDITED BY
Anil Agarwal,
United States Department of Veterans Affairs, United States

REVIEWED BY
Arni S. R. Srinivasa Rao,
Augusta University, United States
Alhaji Cherif,
Merck Sharp & Dohme Corp, United States

*CORRESPONDENCE
Yuedong Wang
yuedong@ucsb.edu

†These authors share first authorship

RECEIVED 04 March 2023
ACCEPTED 28 April 2023
PUBLISHED 02 June 2023

CITATION
Duan J, Li H, Ma X, Zhang H, Lasky R, Monaghan CK, Chaudhuri S, Usvyat LA, Gu M, Guo W, Kotanko P and Wang Y (2023) Predicting SARS-CoV-2 infection among hemodialysis patients using multimodal data.
Front. Nephrol. 3:1179342.
doi: 10.3389/fneph.2023.1179342

Predicting SARS-CoV-2 infection among hemodialysis patients using multimodal data

Juntao Duan[1†], Hanmo Li[1†], Xiaoran Ma[1†], Hanjie Zhang[2], Rachel Lasky[3], Caitlin K. Monaghan[3], Sheetal Chaudhuri[3,4], Len A. Usvyat[3], Mengyang Gu[1], Wensheng Guo[5], Peter Kotanko[2,6] and Yuedong Wang[1*]

[1]Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, United States, [2]Renal Research Institute, New York NY, United States, [3]Fresenius Medical Care, Global Medical Office, Waltham, MA, United States, [4]Division of Nephrology, Maastricht University Medical Center, Maastricht, Netherlands, [5]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia PA, United States, [6]Icahn School of Medicine at Mount Sinai, New York NY, United States

Background: The coronavirus disease 2019 (COVID-19) pandemic has created more devastation among dialysis patients than among the general population. Patient-level prediction models for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection are crucial for the early identification of patients to prevent and mitigate outbreaks within dialysis clinics. As the COVID-19 pandemic evolves, it is unclear whether or not previously built prediction models are still sufficiently effective.

Methods: We developed a machine learning (XGBoost) model to predict during the incubation period a SARS-CoV-2 infection that is subsequently diagnosed after 3 or more days. We used data from multiple sources, including demographic, clinical, treatment, laboratory, and vaccination information from a national network of hemodialysis clinics, socioeconomic information from the Census Bureau, and county-level COVID-19 infection and mortality information from state and local health agencies. We created prediction models and evaluated their performances on a rolling basis to investigate the evolution of prediction power and risk factors.

Result: From April 2020 to August 2020, our machine learning model achieved an area under the receiver operating characteristic curve (AUROC) of 0.75, an improvement of over 0.07 from a previously developed machine learning model published by Kidney360 in 2021. As the pandemic evolved, the prediction performance deteriorated and fluctuated more, with the lowest AUROC of 0.6 in December 2021 and January 2022. Over the whole study period, that is, from April 2020 to February 2022, fixing the false-positive rate at 20%, our model was able to detect 40% of the positive patients. We found that features derived from local infection information reported by the Centers for Disease Control and Prevention (CDC) were the most important predictors, and vaccination status was a useful predictor as well. Whether or not a patient lives in a nursing home was an effective predictor before vaccination, but became less predictive after vaccination.

Conclusion: As found in our study, the dynamics of the prediction model are frequently changing as the pandemic evolves. County-level infection information and vaccination information are crucial for the success of early COVID-19 prediction models. Our results show that the proposed model can effectively identify SARS-CoV-2 infections during the incubation period. Prospective studies are warranted to explore the application of such prediction models in daily clinical practice.

KEYWORDS

COVID-19, hemodialysis, machine learning, prediction, XGBoost

1 Introduction

In December 2019, pneumonia cases of unknown cause emerged in Wuhan, China. Soon the virus was identified as a type of coronavirus named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (1). The resulting acute respiratory disease was named coronavirus disease 2019 (COVID-19). Owing to its highly contagious nature, SARS-CoV-2 soon spread across the globe. As of 3 January 2023, according to the WHO (2), there have been 655,689,115 confirmed cases of COVID-19 (as reported to the WHO) worldwide, including 6,671,624 deaths, equating to a death rate of over 1% among the general population.

Because of older age and multiple comorbidities, dialysis patients are at higher risk of serious complications and death from COVID-19. A greater than 10% case fatality is observed among dialysis patients in different studies (3–5). Considering that patients on maintenance hemodialysis typically have an impaired immune function and are at a higher risk from COVID-19 than the general population, special care is required. Safety procedures have been implemented in dialysis centers (e.g., temperature screenings, universal masking, isolation treatments) to control the spread of SARS-CoV-2 and avoid outbreaks. Specifically, all patients and staff with an elevated body temperature or flu-like symptoms or those who have been exposed to COVID-19 are considered "patients under investigation" (PUI). PUI undergo multiple reverse transcription-polymerase chain reaction (RT-PCR) tests for the detection of SARS-CoV-2 and are treated in dedicated isolation areas (rooms, shifts, or clinics). Although these safety procedures mitigate the rapid spread of SARS-CoV-2 within the dialysis community, they add a significant burden to daily clinic operations.

In the general population, machine learning prediction models have been applied successfully and have reduced economic burden and pandemic control costs (6–9). These COVID-19 prediction models can provide a supportive diagnosis of COVID-19 and prediction of mortality risk and severity using readily available electronic health records (8, 10–12). These efforts add another layer of protection for the general public on top of standard epidemic control procedures, such as social distancing and isolation.

Among the dialysis community, the application of such Artificial Intelligence (AI) supported solutions is still very limited. To alleviate the challenges imposed on daily clinic operations, machine learning prediction models were studied (13, 14). One advantage of adopting these AI models is the possibility of a swift response. AI models can aggregate patient information to detect SARS-CoV-2 infection several days before the RT-PCR test result is available [e.g., 3 days ahead (13)]. The combination of different data sources allows the discovery of features specific to dialysis patients other than general symptoms of COVID-19 (e.g., fever and coughing). Therefore, it may be possible to detect asymptomatic patients during the incubation period.

As the COVID-19 pandemic evolves, it is unclear whether or not previously identified predictors [e.g., residing in a nursing home in (14, 15), clinical and laboratory parameters in (15–17)] are still predictive and previously built machine learning models [e.g., XGBoost in (13)] are still effective for the early detection of COVID-19 cases. Not only has the original virus undergone mutations that have resulted in multiple variants with different clinical presentations (18–20), but also the social environment has significantly changed. For example, lockdowns and social distancing rules have been lifted and vaccination programs have been implemented. Therefore, in contrast to previous studies, we leveraged multiple data sources to study how these changes affected COVID-19 prediction modeling over a much longer period, that is, from January 2020 to February 2022. A longer study period and versatile data sources allowed us to explore the continuous dynamics of COVID-19 prediction and thus provide more reliable and time-tested insights. Ultimately, by combining these insights with AI modeling, we hope to reduce the frequency of false-positive and false-negative predictions, and, consequently, assist dialysis clinics with improving operational efficiency.

2 Materials and methods

2.1 Data collection

Fresenius Kidney Care (FKC) is a large dialysis organization that comprises about 2,400 dialysis clinics in all but one state in the

United States and provides dialysis treatments for approximately one-third of all US dialysis patients. Clinical, treatment, and laboratory information are routinely collected and stored electronically. We identified FKC patients with treatment records from November 2019 to March 2022. Patients suspected of having a SARS-CoV-2 infection at the outpatient dialysis clinics universally underwent RT-PCR testing to diagnose COVID-19. Demographic and socioeconomic information, such as age, dialysis vintage, race, gender, education, employment status, and comorbidities, including hypertension, diabetes, congestive heart failure, and chronic obstructive pulmonary disease, were extracted for each patient. Vaccination information such as the vaccine type (Pfizer, Moderna, or Johnson & Johnson) and administration date were recorded for each patient. Clinical data such as pre- and post-dialysis body temperature, pre- and post-dialysis systolic blood pressure, and interdialytic weight gain (IDWG), treatment data such as treatment time, ultrafiltration volume, ultrafiltration rate, and Kt/V, and intradialytic data, such as intradialytic blood pressure, heart rate, and ultrafiltration rate, were extracted from electronic health records. Laboratory variables such as creatinine, blood urea nitrogen (BUN) and albumin levels, and neutrophil-to-lymphocyte ratio were measured about once a month; hemoglobin levels were measured weekly.

Based on each patient's home zip code, their county-level infection information (including daily new COVID-19 cases and daily COVID-19 deaths) was extracted from the New York Times COVID-19 tracker (21). Other county information, including total population, population density, and percentage of population in poverty, was obtained from the Census Bureau (22). Another feature, "percentage of contracting (PoC) COVID-19", was estimated for each county in the US using a COVID-19 transmission model (23), which represented the daily risk of a susceptible individual contracting COVID-19 in that county.

This study was performed under a protocol reviewed by the Western Institutional Review Board (WIRB; protocol #20212859). WIRB determined that this analysis of deidentified patient data was exempt and did not require informed consent. The analysis was conducted in accordance with the Declaration of Helsinki.

2.2 Confirmed cases and controls

We identified 41,390 COVID-19-positive dialysis patients between 21 January 2020 and 28 February 2022. These positive patients had at least one confirmed positive RT-PCR COVID-19 test during the study period. Only the first confirmed positive date was used in this study. A total of 115,510 negative patients were randomly sampled from all active FKC patients. These COVID-19-negative patients had either a negative or no RT-PCR test during the observation period. Random sampling was performed such that the number of negative patients was approximately three times the total number of positive patients (Figure 1).

We defined a patient's index date as the date of the positive RT-PCR test. For patients who had never reported a positive RT-PCR result, the index date was randomly sampled from the positive patients' index dates.

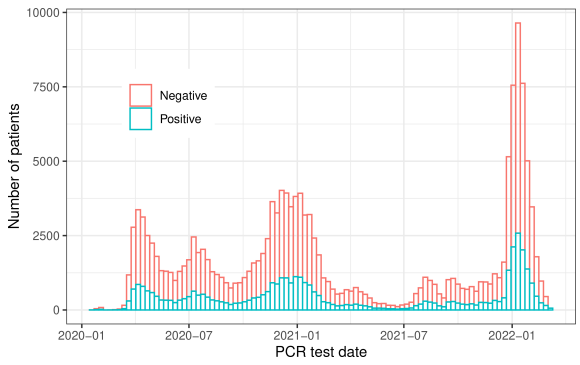

FIGURE 1
Patients' distribution across the whole study period.

We included only patients with (1) at least one hemoglobin test done both 1–14 days and 31–60 days before the individual's prediction date (i.e., 3 days before the index date) and (2) at least one dialysis treatment done both 1–7 days and 31–60 days preceding the prediction date. This was done to ensure that we included only patients who were active, as hemoglobin measurements are done weekly among FKC in-center dialysis patients.

2.3 Data processing and feature engineering

We followed a similar timeline setup to (13). Specifically, we used data from only up to 4 days (see Figure 2) before the index date (expected RT-PCR test date). First, we eliminated outliers from laboratory and treatment measures, which were likely due to manual input errors (e.g., body temperature less than 70°F or greater than 120°F). Second, we created features by taking an average of a variable over two different periods, the normal period and the incubation period. We also used the difference between the mean value in the normal period and the incubation period (the variable name is followed by "_diff"). For treatment and county infection variables, the incubation period was set to 1–7 days before the prediction date. For laboratory measurements, the incubation period was set to 1–14 days before the prediction date due to its less frequent schedule. The normal period is 31–60 days before the prediction date for every variable. Third, since vaccinal immunization decays over time (24), we also calculated the time from the prediction date to the latest vaccination date to reflect this effect. Lastly, the infection or death rate (number of infections or deaths per million people due to COVID-19) at the county level was also calculated to reflect the local epidemic characteristics.

2.4 Models

2.4.1 Training and testing

We selected 116 features, including demographic, treatment, laboratory, and local county variables, up to 4 days before the index

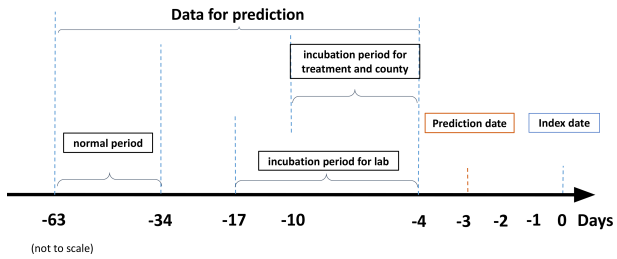

FIGURE 2
Data extraction and prediction timeline.

date. We used these features to predict the risk of a SARS-CoV-2 infection being identified in the next 3 or more days (i.e., on or after the index date). We used a monthly updating strategy to emulate implementation in dialysis clinics. For example, for the prediction in August 2020, we used data before 1 August 2020 as the training sample. Thus the prediction performance on August 2020 was out of the training sample and close to the real-world performance. We compared two types of training as shown in Figure 3, one that used only data within 3 months before the testing period, and another that used all data from the beginning of 2020. The hyperparameters of XGBoost were tuned on the training dataset using cross-validation. The training used binary cross-entropy as the loss function, with weights set as the ratio of positive to negative patients to solve the class imbalance problem.

2.4.2 Evaluation metric

We used the monthly out-of-sample area under the receiver operating characteristic curve (AUROC) to evaluate the performance over the period April 2020 to February 2022. We also calculated the overall AUROC and precision–recall curve (PRC) with aggregated monthly predictions.

2.4.3 Feature importance

We used SHapley Additive exPlanations (SHAP) values to identify the influential variables on monthly testing predictions (25, 26). For each specific prediction, the SHAP value was computed for every variable, which measures how much the predicted value is affected by each variable used in the XGBoost model. The overall feature importance of each variable can be quantified by the mean absolute value of SHAP values for each variable across all observations.

FIGURE 3
Two types of training strategy: 3-month rolling and cumulative.

3 Results

3.1 Model performance

To assess the impact of different training strategies on model performance, we compared overall AUROC and PRC in Figure 4, and monthly AUROC in Figure 5.

Overall, testing performance in Figure 4 was calculated by aggregating all monthly predictions. For instance, overall precision was computed as the ratio of correctly predicted positive COVID-19 cases to the total number of positive COVID-19 cases over the entire testing period, that is, from April 2020 to February 2022. As shown in Figure 4, there is minimal difference between the 3-month rolling and cumulative models. In both cases, with the false-positive rate fixed at 20%, the true-positive rate is slightly above 40%.

In Figure 5, we find the two training strategies exhibit only slight variations in their monthly performance. The 3-month rolling models are more responsive to recent changes, such as sudden waves. In Figure 5, the AUROC started at around 0.75 and dropped slightly to 0.70 on August 2020. After that, the performance fluctuates between 0.60 and 0.70. Multiple reasons may cause this trend of performance degradation. First, predictive features may become unproductive over time for various reasons, which will be discussed later when investigating feature importance. Second, as the reopening policy was rolled out, pandemic characteristics have changed. Third, data quality is degrading over time due to under-reporting because of asymptomatic cases and less frequent updates of county infection reports by CDC[1].

In the study of Monaghan et al. (13), FKC patients' data from 27 February 2020 to 8 September 2020 was used to build a machine learning model for early prediction of COVID-19 cases. Their model focused on biological changes in clinical biomarkers and achieved a testing AUROC of 0.68. Compared with Monaghan et al. (13), we achieved a higher testing AUROC at around 0.75 (Figure 5) before 31 August 2020. The performance improvement is due to additional data being included; local county infection information in particular played an important role. After 2021, as vaccination was rolled out, we identified that vaccination become a crucial predictor, which only our study was able to investigate..

3.2 Feature importance

After ranking the mean absolute value of SHAP values, the top 40 features were identified and summarized. They are shown in Tables 1, 2. The top nine features are shown in Figure 6.

1 CDC COVID-19 surveillance switched to the weekly report on October 20, 2022

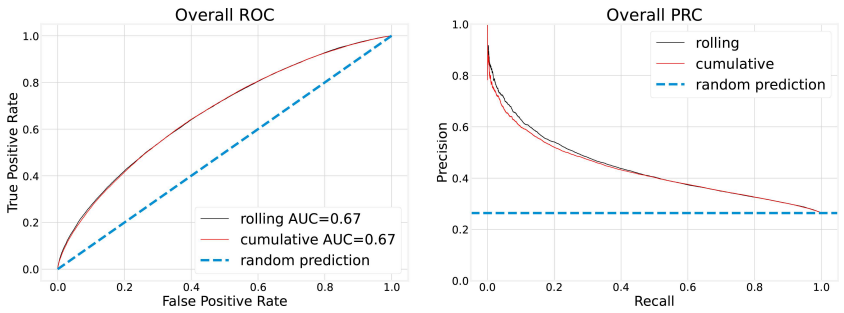

FIGURE 4
Overall testing performance is calculated with aggregated monthly predictions.

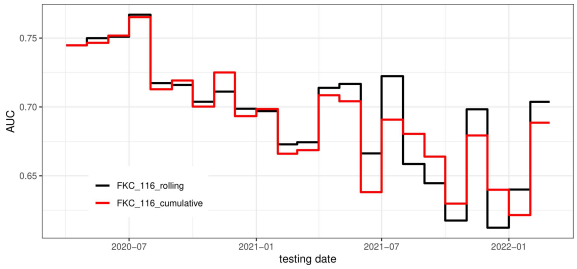

FIGURE 5
Monthly AUROC.

The top two features that remained important throughout the whole time period were the average number of new COVID-19 cases per million population in the incubation period ("covid_new_cases_incubation_rate") and the difference in the average number of new COVID-19 cases per million population between the incubation period and the normal period (covid_new_cases_diff_rate). In Figure 7, "covid_new_cases_incubation_rate" and "covid_new_-cases_diff_rate" have a positive correlation with COVID-19 cases. These two features were derived from local county COVID-19 cases reports. Another important feature of local information is population density ("population_density"), as high population densities can increase the risk of spreading SARS-CoV-2. In areas with high population densities, such as cities or densely populated neighborhoods, it can be difficult to maintain physical distancing and limit close contact with others.

TABLE 1 Demographics and categorical features of hemodialysis patients with and without a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection.

Variable	Unaffected patients	COVID-19-positive patients
Number of patients on HD	115,510	41,503
Male, n (%)	67,717 (59)	22,857 (55)
Hispanic or Latino, n (%)	15,748 (14)	7,625 (19)
Diabetes, n (%)	48,928 (45)	19,256 (49)
Nursing home, n (%)	6,025 (6)	5,290 (15)
Race, *n* (%)		
American Indian or Alaska Native	877 (1)	587 (1)
Asian	4,063 (4)	988 (2)
Black people or African American	38,664 (35)	13,632 (34)
Native Hawaiian or Other Pacific Islander	1,285 (1)	481 (1)

(Continued)

TABLE 1 Continued

Variable	Unaffected patients	COVID-19-positive patients
White people	65,717 (59)	24,434 (61)
Education, n (%)		
8 or less years of school	8,592 (7)	4,289 (10)
Current student	78 (0)	31 (0)
GED	2,951 (3)	1,142 (3)
Graduated from 2- or 4-year college	17,563 (15)	4,871 (12)
Graduated high school	44,257 (38)	16,113 (39)
Graduate school	4,960 (4)	1,135 (3)
More than 8 years but less than 12 years	14,252 (12)	6,390 (15)
Some college	18,677 (16)	6,156 (15)
Vocational/technical school	4,074 (4)	1,364 (3)
Other	106 (0)	12 (0)

COVID-19, coronavirus disease 2019; GED, general educational development.

TABLE 2 Numerical features in top 40 most important features of hemodialysis patients with and without a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection.

Variable	Unaffected patients, mean ± SD	COVID-19-positive patients, mean ± SD
Age (years)	64.1 ± 4.15	62.84 ± 4.32
Height (cm)	168.96 ± 1.37	168.21 ± 1.38
BMI (kg/m^2)	28.82 ± 7.49	29.77 ± 7.86
Number of days since last vaccination	123.46 ± 99.16	144.13 ± 104.49
Dialysis vintage (days)	1,481.85 ± 501.48	1,530.17 ± 466.68
County-level local information per million population		
Daily infected COVID cases	642.25 ± 863.73	755.07 ± 868.42
Change in daily infected COVID cases	380.69 ± 872.24	487.99 ± 883.68
Change in daily COVID death	186.15 ± 98.99	197.52 ± 76.73
Treatment information		
Pre-HD body temperature (°F)	97.36 ± 0.61	97.45 ± 0.64
Post-HD body temperature (°F)	97.45 ± 0.52	97.51 ± 0.56
Change in post-HD body temperature (°F)	−0.01 ± 0.45	0.05 ± 0.49
Change in pre-HD weight loss (kg)	−0.21 ± 2.42	−0.4 ± 2.51
Change in weight (kg)	0.04 ± 13.54	0.04 ± 15.42
Change in IDWG (kg)	0.02 ± 0.87	−0.11 ± 0.96
Change in pre-HD pulse (BPM)	0.04 ± 7.46	0.76 ± 7.51
Change in Post-HD pulse (BPM)	−0.05 ± 7.34	0.98 ± 7.60
Change in max-HD pulse (BPM)	0.07 ± 8.05	1.07 ± 8.34
Min-HD pulse (BPM)	65.8 ± 10.52	66.8 ± 10.60

(Continued)

TABLE 2 Continued

Variable	Unaffected patients, mean ± SD	COVID-19-positive patients, mean ± SD
Max-HD sitting SBP (mmHg)	155.58 ± 22.68	158.13 ± 22.89
Post-HD sitting SBP (mmHg)	139.33 ± 21.03	141.4 ± 21.38
Change in pre-HD sitting SBP (mmHg)	0.19 ± 15.69	−0.87 ± 16.56
Change in pre-HD sitting DBP (mmHg)	0.1 ± 9.07	−0.43 ± 9.43
Laboratory measurements		
Albumin (g/dL)	3.82 ± 0.42	3.73 ± 0.44
Calcium (mg/dL)	8.94 ± 0.68	8.85 ± 0.69
Change in % of monocytes	0 ± 1.24	0.15 ± 1.42
Change in WBC count (10^{10}/L)	−0.05 ± 2.92	−0.14 ± 2.05
Hgb (g/dL)	10.73 ± 1.25	10.64 ± 1.23
TSAT (%)	32.63 ± 14	31.7 ± 14.18
URR	74.58 ± 6.63	74.82 ± 6.34
WBC count (10^{10}/L)	6.96 ± 3.59	6.85 ± 3.00
% of eosinophils	4.33 ± 2.66	4.09 ± 2.62

COVID-19, coronavirus disease 2019; BMI, body mass index; HD, hemodialysis; IDWG, interdialytic weight gain; TSAT, transferrin saturation; URR, urea reduction ratio.

Whether or not a patient lives in a nursing home ("nursing_home") was a robust top predictor before 2021, but its significance has gradually declined. Following the implementation of vaccination in 2021, a substantial decrease was observed. As shown in Figure 7, "nursing_home" was ranked as the second most significant factor that positively correlated with COVID-19 cases in July 2020. However, by July 2021 it had dropped out of the top five features.

Clinical (treatment) and demographic information such as body mass index ("BMI"), interdialytic weight gain between the incubation period and normal period ("IDWG_diff"), and vintage appeared among the top nine features, which confirms the findings in Chaudhuri et al. (16) that clinical and laboratory variables are predictive. As shown in Figure 7, pre-dialysis body temperature during the incubation period ("pre_temperature_incubation") was

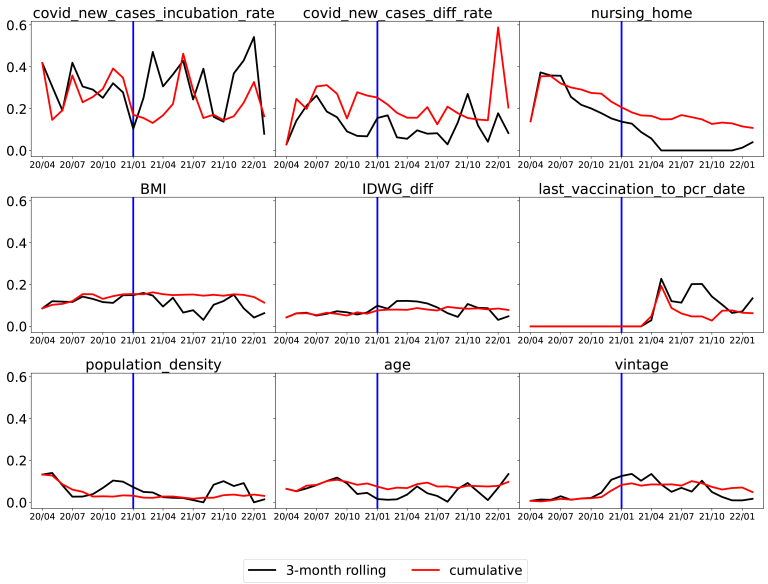

FIGURE 6
Monthly average absolute SHAP value for the top nine important features. For each feature, the black line is for the 3-month rolling model, and the red line is for the cumulative model. The vertical blue line at the end of 2020 is a separation of whether or not vaccination is available (U.S. HHS, Vaccination in the US began on 14 December 2020). (Note that the x-axis is the date in "yy/mm" format).

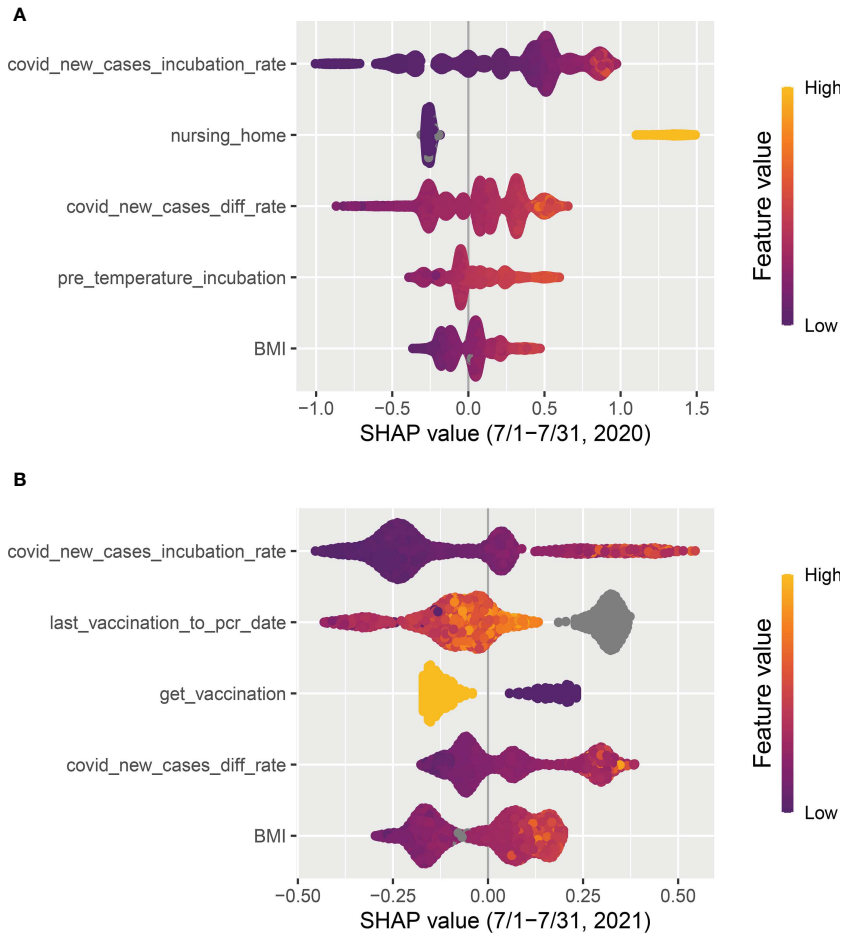

FIGURE 7
Top five features ranked by absolute SHAP value. **(A)** The model was trained with the data from 1 April to 30 June in 2020 and tested on data from 1 July to 31 July, 2020. **(B)** The model was trained with data from 1 September to 30 November in 2021 and tested on data from 1 December to 31 December, 2021.

also identified as one of the top five predictive features in July 2020 and had a positive correlation with COVID-19 cases. However, similar to "nursing_home", it had dropped out of the top five features by July 2021.

The variable "last_vaccination_to_pcr_date" is defined as the difference between a patient's prediction date and the latest vaccination date. As illustrated in Figure 7, "last_vaccination_to_pcr_date" emerged as a significant predictor in the middle of 2021, ranking first in importance, and had a positive correlation with COVID-19 cases. Specifically, a missing or large value in "last_vaccination_to_pcr_date" implied that the patient had a higher chance of being identified as a positive COVID-19 case. However, by the end of 2021 its significance was reduced Figure 6, possibly due to a decrease in the efficacy of vaccination-induced antibodies over time. Similarly, another vaccination-related feature, "get_vaccination", which was defined as whether or not a patient had received vaccines before the prediction date, has a negative correlation with positive COVID-19 cases, as shown in Figure 7 compare before after vaccination.

Comparing the two training strategies, the 3-month rolling models (represented by black lines) generally produced rougher curves for SHAP values, as they were quicker to respond to rapid changes. In contrast, cumulative models (represented by red lines) utilized accumulated data, making them less responsive to changes such as the introduction of vaccination.

4 Discussion

We have successfully developed a machine learning model that utilizes multiple data sources to detect early COVID-19 infections in maintenance hemodialysis patients. We demonstrated that the proposed machine learning model achieved clinically meaningful performance by monthly testing throughout the COVID-19 pandemic. Overall, the model was able to identify 40% of COVID-19 patients (with a 20% false-positive rate) before they were identified by an RT-PCR COVID-19 test. This can significantly aid dialysis clinics in preventing the spread of the virus by implementing targeted procedures for identified patients.

More importantly, apart from the patient's laboratory and treatment information (e.g., body temperature) used by Monaghan et al. (13), we identified two other sources of

information that are more critical for such prediction models. The first is local county infection data provided by CDC. County-level COVID-19 infection information reflects how likely SARS-CoV-2 is to spread within the patient's community. The second is the patient's vaccination record, which reflects how likely SARS-CoV-2 will infect a patient after contact. As the dynamics of COVID-19 change, previously important features, such as whether or not a patient lives in a nursing home, become less predictive.

There are limitations to our study. First, the positive COVID-19 diagnosis labels are limited to the positive patients' specific RT-PCR test date. As the COVID-19 pandemic is a continuous and dynamic process, there are dates before and after the RT-PCR test date that should also be annotated as positive labels. However, without enough information, it is difficult to select these periods. Furthermore, the timing of the RT-PCR test relative to infection can vary based on symptom presentation and other factors, adding additional variability to the RT-PCR test dates. It is worth further investigating the cutoff dates to further improve COVID-19 early detection. Second, as limited by anticipated data availability and model integration into clinical systems, the prediction date was set to 3 days before the index date. Ideally, with real-time data aggregation, one could set the prediction date to the index date. This will likely improve the detection performance as the most recent laboratory and treatment data can be used. In addition, more advanced methods such as deep learning could potentially further improve the accuracy of COVID-19 detection (27–30). This avenue will be explored in future research endeavors.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding author.

Ethics statement

The studies involving human participants were reviewed and approved by the Western Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

YW, PK, HZ, WG, MG, and LU conceived and designed the study. RL and SC collected the data. JD, HL, and XM performed the statistical analysis and wrote the manuscript's original draft. CM provided advice on the prediction model published by Kidney360 in 2021. All authors contributed to the article and approved the submitted version.

Funding

Research reported in this publication was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under award number R01DK130067.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors PK and HZ declared that they were an editorial board member of Frontiers at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Author disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Gorbalenya AE, Baker SC, Baric RS, de Groot RJ, Drosten C, Gulyaeva AA, et al. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2. *Nat Microbiol* (2020) 5(4):536–44. doi: 10.1038/s41564-020-0695-z

2. World Health Organization. *Covid-19 dashboard*. Available at: https://covid19.who.int/.

3. Salerno S, Messana JM, Gremel GW, Dahlerus C, Hirth RA, Han P, et al. Covid-19 risk factors and mortality outcomes among medicare patients receiving long-term dialysis. *JAMA Netw Open* (2021) 4(11):e2135379–e2135379. doi: 10.1001/jamanetworkopen.2021.35379

4. Yavuz D, Karagöz Özen DS, Demirağ MD. Covid-19: mortality rates of patients on hemodialysis and peritoneal dialysis. *Int Urol Nephrol* (2022), 1–6. doi: 10.1007/s11255-022-03193-6

5. Appelman B, Oppelaar JJ, Broeders L, Wiersinga WJ, Peters-Sengers H, Vogt L. Mortality and readmission rates among hospitalized covid-19 patients with varying stages of chronic kidney disease: a multicenter retrospective cohort. *Sci Rep* (2022) 12 (1):1–8. doi: 10.1038/s41598-022-06276-7

6. Zoabi Y, Deri-Rozov S, Shomron N. Machine learning-based prediction of covid-19 diagnosis based on symptoms. *NPJ Digital Med* (2021) 4(1):1–5. doi: 10.1038/s41746-020-00372-6

7. Mardian Y, Kosasih H, Karyana M, Neal A, Lau CY. Review of current covid-19 diagnostics and opportunities for further development. *Front Med* (2021) 8:615099. doi: 10.3389/fmed.2021.615099

8. Alballa N, Al-Turaiki I. Machine learning approaches in covid-19 diagnosis, mortality, and severity risk prediction: a review. *Inf Med Unlocked* (2021) 24:100564. doi: 10.1016/j.imu.2021.100564

9. Ayris D, Imtiaz M, Horbury K, Williams B, Blackney M, Hui See CS, et al. Novel deep learning approach to model and predict the spread of covid-19. *Intelligent Syst Appl* (2022) 14:200068. doi: 10.1016/j.iswa.2022.200068

10. Kukar M, Gunčar G, Vovko T, Podnar S, Černelč P, Brvar M, et al. Covid-19 diagnosis by routine blood tests using machine learning. *Sci Rep* (2021) 11(1):1–9. doi: 10.1038/s41598-021-90265-9

11. Mamidi TKK, Tran-Nguyen TK, Melvin RL, Worthey EA. Development of an individualized risk prediction model for covid-19 using electronic health record data. *Front Big Data* (2021) 4:675882. doi: 10.3389/fdata.2021.675882

12. Bayat V, Phelps S, Ryono R, Lee C, Parekh H, Mewton J, et al. A severe acute respiratory syndrome coronavirus 2 (sars-cov-2) prediction model from standard laboratory tests. *Clin I Dis* (2021) 73(9):e2901–7. doi: 10.1093/cid/ciaa1175

13. Monaghan CK, Larkin JW, Chaudhuri S, Han H, Jiao Y, Bermudez KM, et al. Machine learning for prediction of patients on hemodialysis with an undetected sars-cov-2 infection. *Kidney* (2021) 360:456. doi: 10.34067/KID.0003802020

14. Haarhaus M, Santos C, Haase M, Mota Veiga P, Lucas C, Macario F. Risk prediction of covid-19 incidence and mortality in a large multi-national hemodialysis cohort: implications for management of the pandemic in outpatient hemodialysis settings. *Clin Kidney J* (2021) 14(3):805–13. doi: 10.1093/ckj/sfab037

15. Hsu CM, Weiner DE, Aweh G, Miskulin DC, Manley HJ, Stewart C, et al. Covid-19 among us dialysis patients: risk factors and outcomes from a national dialysis provider. *Am J Kidney Dis* (2021) 77(5):748–56. doi: 10.1053/j.ajkd.2021.01.003

16. Chaudhuri S, Lasky R, Jiao Y, Larkin J, Monaghan C, Winter A, et al. Trajectories of clinical and laboratory characteristics associated with covid-19 in hemodialysis patients by survival. *Hemodialysis Int* (2022) 26(1):94–107. doi: 10.1111/hdi.12977

17. Kooman J, Carioni P, Kovarova V, Arkossy O, Winter A, Zhang Y, et al. Modifiable risk factors are important predictors of covid-19-related mortality in patients on hemodialysis. *Front Nephrol* (2022) 2:907959. doi: 10.3389/fneph

18. Puenpa J, Rattanakomol P, Saengdao N, Chansaenroj J, Yorsaeng R, Suwannakarn K, et al. Molecular characterisation and tracking of severe acute respiratory syndrome coronavirus 2 in thailand, 2020–2022. *Arch Virol* (2023) 168 (1):26. doi: 10.1007/s00705-022-05666-6

19. Bouzid D, Visseaux B, Kassasseya C, Daoud A, Fémy F, Hermand C, et al. Comparison of patients infected with delta versus omicron covid-19 variants presenting to paris emergency departments: a retrospective cohort study. *Ann Internal Med* (2022) 175(6):831–7. doi: 10.7326/M22-0308

20. Araf Y, Akter F, Yd T, Fatemi R, Parvez MSA, Zheng C, et al. Omicron variant of sars-cov-2: genomics, transmissibility, and responses to current covid-19 vaccines. *J Med Virol* (2022) 94(5):1825–32. doi: 10.1002/jmv.27588

21. The New York Times. *Coronavirus (Covid-19) data in the united states* (2021). Available at: https://github.com/nytimes/covid-19-data.

22. Census Bureau, United States. Available at: https://www.census.gov/data/developers/data-sets.html.

23. Li H, Gu M. Robust estimation of sars-cov-2 epidemic in us counties. *Sci Rep* (2021) 11(1):1–16. doi: 10.1038/s41598-021-90195-6

24. Wang X, Han M, Fuentes LR, Thwin O, Grobe N, Wang K, et al. Sars-cov-2 neutralizing antibody response after three doses of mrna1273 vaccine and covid-19 in hemodialysis patients. *Front Nephrol* (2022). doi: 10.3389/fneph.2022.926635

25. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst* (2017) 4768–77. doi: 10.5555/3295222.3295230

26. Shapley LS. *A value for n-person games, contributions to the theory of games II. annals of mathematics studies*. Princeton: Princeton University Press (1953).

27. Srinivasa Rao ASR, Vazquez JA. Identification of COVID-19 can be quicker through artificial intelligence framework using a mobile phone–based survey when cities and towns are under quarantine. *Infect Control Hosp Epidemiol* (2020) 41(7):826–30. doi: 10.1017/ice.2020.61

28. Liu Y, Zhou Y, Liu X, Dong F, Wang C, Wang Z. Wasserstein GAN-based small-sample augmentation for new-generation artificial intelligence: a case study of cancer-staging data in biology. *Engineering* (2019) 5(1):156–63. doi: 10.1016/j.eng.2018.11.018

29. Tasci E, Zhuge Y, Camphausen K, Krauze AV. Bias and class imbalance in oncologic data–towards inclusive and transferrable AI in Large scale oncology data sets. *Cancers* (2022) 14(12):2897. doi: 10.3390/cancers14122897

30. Acs B, Rantalainen M, Hartman J. Artificial intelligence as the next step towards precision pathology. *J Internal Med* (2020) 288(1):62–81. doi: 10.1111/joim.13030