PLOS ONE

RESEARCH ARTICLE

Predicting clinical outcomes of SARS-CoV-2 infection during the Omicron wave using machine learning

Steven Cogill[1,2], **Shriram Nallamshetty**[1], **Natalie Fullenkamp**[1], **Kent Heberer**[1,2], **Julie Lynch**[3,4], **Kyung Min Lee**[3], **Mihaela Aslan**[5,6], **Mei-Chiung Shih**[1,7], **Jennifer S. Lee**[1,2,8]*

1 VA Palo Alto Cooperative Studies Program Coordinating Center, Palo Alto, CA, United States of America, **2** Big Data-Scientist Training Enhancement Program at VA Palo Alto Health Care System, Palo Alto, CA, United States of America, **3** VA Informatics and Computing Infrastructure, VA Salt Lake City Health Care System, Salt Lake City, UT, United States of America, **4** Division of Epidemiology, Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT, United States of America, **5** VA Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West Haven, CT, United States of America, **6** Department of Medicine, Yale University School of Medicine, New Haven, CT, United States of America, **7** Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, United States of America, **8** Division of Endocrinology, Department of Medicine, Gerontology, and Metabolism, and by Courtesy, of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, United States of America

* Jennifer.Lee23@va.gov

OPEN ACCESS

Citation: Cogill S, Nallamshetty S, Fullenkamp N, Heberer K, Lynch J, Lee KM, et al. (2024) Predicting clinical outcomes of SARS-CoV-2 infection during the Omicron wave using machine learning. PLoS ONE 19(4): e0290221. https://doi.org/10.1371/journal.pone.0290221

Editor: Omid Dadras, University of Bergen: Universitetet i Bergen, NORWAY

Received: August 4, 2023

Accepted: February 20, 2024

Published: April 25, 2024

Data Availability Statement: Patient related and note data cannot be shared publicly because it involves sensitive human subject data. Data may be available for researchers who meet the criteria for access to confidential data after evaluation from VA Research and Development Committees. As a VA national legal policy (VHA Directive 1605.01), VA will only share patient data if there is a fully executed contract in place for the specific project. Data requests may be sent to: VA Information Resource Center (VIReC) Building 18 Hines VA

Abstract

The Omicron SARS-CoV-2 variant continues to strain healthcare systems. Developing tools that facilitate the identification of patients at highest risk of adverse outcomes is a priority. The study objectives are to develop population-scale predictive models that: 1) identify predictors of adverse outcomes with Omicron surge SARS-CoV-2 infections, and 2) predict the impact of prioritized vaccination of high-risk groups for said outcome. We prepared a retrospective longitudinal observational study of a national cohort of 172,814 patients in the U.S. Veteran Health Administration who tested positive for SARS-CoV-2 from January 15 to August 15, 2022. We utilized sociodemographic characteristics, comorbidities, and vaccination status, at time of testing positive for SARS-CoV-2 to predict hospitalization, escalation of care (high-flow oxygen, mechanical ventilation, vasopressor use, dialysis, or extracorporeal membrane oxygenation), and death within 30 days. Machine learning models demonstrated that advanced age, high comorbidity burden, lower body mass index, unvaccinated status, and oral anticoagulant use were the important predictors of hospitalization and escalation of care. Similar factors predicted death. However, anticoagulant use did not predict mortality risk. The all-cause death model showed the highest discrimination (Area Under the Curve (AUC) = 0.903, 95% Confidence Interval (CI): 0.895, 0.911) followed by hospitalization (AUC = 0.822, CI: 0.818, 0.826), then escalation of care (AUC = 0.793, CI: 0.784, 0.805). Assuming a vaccine efficacy range of 70.8 to 78.7%, our simulations projected that targeted prevention in the highest risk group may have reduced 30-day hospitalization and death in more than 2 of 5 unvaccinated patients.

Hospital (151V) 5000 S. 5th Avenue Hines, IL 60141-3030 (Email: virec@va.gov).

Funding: The work was supported by the VA Cooperative Studies Program (all authors) and the VA Palo Alto Healthcare System (S. Nallamshetty, N. Fullenkamp, J. Lee). 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.

Introduction

The World Health Organization (WHO) estimates that the COVID-19 pandemic has resulted in over 521 million infections and 6.2 million deaths globally [1]. High mutation rates and the relatively rapid emergence of SARS-CoV-2 variants led to multiple surges that have strained healthcare systems worldwide. The Omicron (B.1.1.529) variant became the predominant cause of SARS-CoV-2 infections in the U.S. by January 2022 [2, 3], after identification in South Africa in November 2021 [4, 5]. Although Omicron variants and sub-variants have been linked to lower rates of hospitalization and death, [3, 6–8] Omicron-driven surges continued to challenge healthcare systems due to higher infectivity, partial vaccine escape, and antibody resistance [3, 7].

Predictive modeling during the pandemic has provided crucial insight into clinical outcomes with COVID-19 infections; however, to date, these risk prediction tools have largely not included data for Omicron variants and have inconsistently incorporated important clinical factors such as vaccination status [9–12]. In this study, we first applied machine learning (ML) models to identify baseline patient characteristics that predict risk for hospitalization, escalation of care, and mortality among SARS-CoV-2 positive US Veterans during a recent seven-month observation period (January 15 –August 15, 2022) when Omicron variants predominated. Our models incorporated previously under-utilized factors including vaccination status. Then, we extended our models to quantify the predicted impact of a mitigating strategy such as prioritized vaccination of high-risk groups on reducing the short-term risk of hospitalization, escalation of care, and death during the observation period. To do this, we utilized a well-characterized cohort of U.S. Veterans with SARS-CoV-2 infection in a national Veteran Health Administration (VHA) database.

Materials and methods

Study cohort

Our study cohort consisted of all 172,814 Veterans who first tested positive for COVID-19 between January 15 and August 15, 2022, as captured by the VHA's COVID-19 Shared Data Resource with data curation within the VHA's Corporate Data Warehouse (CDW). No new data were collected, and no direct patient (or participant) contact took place. Patients' curated electronic health records in the VHA's CDW were analyzed behind the VHA secured firewall as part of the VHA research data initiative, Leveraging Electronic Health Information to Advance Precision medicine (LEAP, CSP#2012), which has been approved by VHA's Central Institutional Review Board and Research & Development Committees at 3 VA Medical Centers (Salt Lake City, Palo Alto, and West Haven). The VHA's CIRB approved a waiver of requirement to obtain informed consent. The date of the first positive test is defined as the index date. For the selected cohort within the data resource, there were no missing data for the selected fields and unknown covariates were indicated as such. Patients outside the age range of 18 to 100, outside the Body Mass Index (BMI) range of 15 to 100, or who experienced reinfection during the 8-month observation period were excluded from the analysis.

Study outcomes

We predicted the risk of developing one of the following three distinct, non-mutually exclusive clinical outcomes representing SARS-CoV-2 severity within 30 days of infection: (i) hospitalization, (ii) escalation of care (defined as the need for high-flow supplemental oxygen, mechanical ventilation, vasopressors, renal replacement therapy [with no prior dialysis in the preceding two years], or extracorporeal membrane oxygenation [ECMO]), and (iii) all-cause mortality. Patients who tested positive for SARS-CoV-2 were deemed to have 'mild' infection if they did not experience any of the three outcomes of interest within 30 days of infection. The Upset plot was generated using the UpsetR package [13].

Clinical features

A total of 159 patient characteristics including medical comorbidities, demographic data, vaccination status, and comorbidity indices were available for each patient prior to feature selection. The medical history included pre-existing conditions, procedures, and medications. All medical history values were classified using a Boolean system for presence or absence of the specific medical condition within two years prior to the current COVID-19 infection. Demographic and clinical data employed in the modeling included age, sex, race/ethnicity, blood type, BMI, veteran status, whether overweight at index date, rurality of current residence, and veteran priority status (a surrogate for income status and benefits eligibility). These covariates were multimodal (float, categorical and Boolean). Vaccination status was represented as a categorical score from 0 to 5 as follows: 0 = no vaccination, 1 = partial-mRNA vaccination, 2 = full vaccination (two doses of mRNA or a single dose of viral vector-based vaccine) > 5 months from index date, 3 = fully-vaccinated and boosted >5 months prior to the index date, 4 = fully-vaccinated <5 months prior to the index date, 5 = fully-vaccinated and boosted <5 months prior to the index date. Vaccines given outside of the VHA were available in the VHA COVID-19 Shared Data Resource and reflected in our dataset. Vaccination status accounted for a two-week efficacy window. Medical comorbidity burden was assessed by Charlson Comorbidity Index (CCI) [14] and Elixhauser Index [15] scores for the two years prior to infection. An overall CCI and Elixhauser index score was also determined. A complete list of covariates is included in S1 Table.

Model development and performance

For each of the 3 main outcomes of interest, we developed a distinct binary model that incorporated 159 unique covariate features using gradient boosting automated machine learning methods. A recursive feature elimination approach was used to find the most parsimonious models. Our data was split chronologically with training/validation data from January 15, 2022 to April 15, 2022 and our test data from April 16, 2022 to August 15, 2022. Covariates with variance lower than 1% within the training set were removed, and non-binary values were scaled from 0 to 1.

Model training and optimization were performed on the training and validation sets. The H2O AI package for automated machine learning was used to train each model and the validation set was used for benchmarking the optimization process [16]. An initial heuristic search through available modeling methods using this package identified gradient boosting machines as the highest performers (S2 Table). Stacked models were not considered due to low interpretability to performance tradeoff. All subsequent modeling was done using gradient boosting machines. Class imbalance within this study is a bias towards patients not having a severity outcome, and this was overcome by oversampling of the minority class where patients did have a severity outcome in training of the models to allow for higher predictive performance. The binary threshold for the models was calculated by finding the threshold with the max geometric mean for specificity and sensitivity on the test set. The 95% confidence intervals for the performance metrics were determined using the stat_util python package and its bootstrapping method with 100 iterations [17].

All reported performance metrics were generated on the set aside test set. Receiver operator characteristic (ROC) and precision recall curves and their respective area under curve (AUC) were calculated using the scikit-learn metrics package [18]. The precision recall curves were normalized by using sample weights.

Model interpretation and applications

Feature importance values were extracted from the H2O generated models [16]. Relative importance is calculated as the decrease in mean squared error weighted by the number of samples passing through a given node for all trees. The percentage reported here is the fraction of a given feature against all other feature relative importance values.

Shapley Additive exPlanations (SHAP) values were generated on the test set using the SHAP python package and a tree-based explainer [19]. SHAP values were calculated on random sampling of 1,000 patients from the test set. Summary plots were generated by plotting the SHAP values in a bee swarm fashion.

For simulating the impact of targeted vaccinations, we selected the unvaccinated subset of our cohort from our test set. For each strategy scenario, we projected the potential reduction in outcomes if the patients were fully vaccinated (4 score in our vaccination status). The projection required two steps. The first was to project how many symptomatic infections would be prevented and thus prevent the outcome. To accomplish this, we randomly sampled and removed patients from our target group based on a published vaccine efficacy 95% CI range of 0.708 to 0.787 which we sampled from in a uniform fashion [20]. The second was to project for the remaining patients in our target group whether being fully vaccinated would have prevented the outcome. For this we used our model and determined if their predicted outcome changed when we altered the vaccination status score from 0 to 4. We then summed the remaining outcomes in our target group to determine the reduction. The 95% confidence intervals for the projections were determined using the stat_util python package and its bootstrapping method with 100 iterations [17].

Results

Patient population and clinical predictors of COVID-19 infection severity

In a national VHA cohort of 172,814 Veterans who first tested positive for SARS-CoV-2 during a period in which the Omicron variant predominated (January 15-August 15, 2022), the median age was 62 years and 84% were men (Table 1). The racial/ethnic composition of the cohort was typical for a US Veteran population; 65.5% of the patients were white, 19% were black, and 9.4% were Hispanic. Asian, Native Hawaiian or Pacific Islander, and American Indian or Alaskan Native Veterans each represented approximately 1% of the cohort. (Table 1).

Baseline characteristics of study cohort of U.S. Veterans who tested positive for SARS-CoV-2.

Overall, 89.5% of Veterans had mild SARS-CoV-2 infections. Among Veterans who tested positive for SARS-CoV-2, 9.2% required hospitalization, 2.2% needed escalation of care, and 1.5% died (Table 1 and Fig 1). In the subset of hospitalized infected patients, a higher percentage required escalation of care (18%) and died (7%) compared to the overall cohort (Fig 1). Patients who died or required hospitalization and/or escalation of care were older and more likely to be male. Conversely, patients who had mild infections had a higher body mass index (BMI) than those who did not (Table 1). A higher percentage of patients who died were white, compared to the overall cohort (78.1% vs 65.5%). In contrast, a lower percentage of patients who died were black, compared to those in the overall cohort (11.3% vs. 19%) (Table 1).

Patients with non-mild infections had higher prevalence of diabetes, congestive heart failure, cerebrovascular disease, chronic kidney disease, and cirrhosis. Dementia was also more prevalent among patients who required hospitalization, required escalation of care, or died within 30 days after testing positive. While chronic lung disease also was more prevalent, Table 1. 30-day outcomes after a positive SARS-CoV-2 test.

	Mild	Hospitalized	Escalation	Mortality	Overall
Characteristics	**n = 154,740 (89.5%)**	**n = 15,831 (9.2%)**	**n = 3,723 (2.2%)**	**n = 2,578 (1.5%)**	**n = 172,814**
Age (median [IQR])	61 [47, 72]	72 [64, 78]	72 [64, 77]	77 [72, 85]	62 [49, 73]
BMI, mean (SD)	30.3 (6.2)	28.1 (6.9)	28.4 (7.0)	26.7 (6.7)	30.1 (6.3)
Men, No. (%)	127,997 (82.7)	14,962 (94.5)	3,515 (94.4)	2,529 (98.1)	145,093 (84.0)
Race, No. (%)					
White	100,477 (64.9)	10,941 (69.1)	2,706 (72.7)	2,013 (78.1)	113,156 (65.5)
Black	29,132 (18.8)	3,379 (21.3)	640 (17.2)	291 (11.3)	32,762 (19.0)
Asian	2,514 (1.6)	94 (0.6)	26 (0.7)	16 (0.6)	2,626 (1.5)
Native American/Alaska Native	1,274 (0.8)	133 (0.8)	30 (0.8)	21 (0.8)	1,423 (0.8)
Native Hawaiian/Other Pacific Islander	1,694 (1.1)	104 (0.7)	40 (1.1)	24 (0.9)	1,825 (1.1)
Unknown	19,649 (12.7)	1,180 (7.5)	281 (7.5)	213 (8.3)	21,022 (12.2)
Hispanic or Latino, No. (%)	14787 (9.6)	1254 (7.9)	319 (8.6)	166 (6.4)	16184 (9.4)
Vaccination Status, No (%)					
0-Unvaccinated	44,893 (29.0)	4,500 (28.4)	1,211 (32.5)	1,067 (41.4)	50,263 (29.1)
1-Parital mRNA (1 dose)	2,641 (1.7)	333 (2.1)	70 (1.9)	60 (2.3)	3,018 (1.7)
2-Fully Vaccinated > 5 months prior	38,675 (25.0)	4099 (25.9)	948 (25.5)	725 (28.1)	43,376 (25.1)
3-Fully Vaccinated with Booster > 5 months prior	42,526 (27.5)	4,263 (26.9)	899 (24.1)	387 (15.0)	47,195 (27.3)
4-Fully Vaccinated < = 5 months prior	3,256 (2.1)	305 (1.9)	68 (1.8)	36 (1.4)	3,598 (2.1)
5-Fully Vaccinated with Booster < = 5 months prior	22,749 (14.7)	2,331 (14.7)	527 (14.2)	303 (11.8)	25,364 (14.7)
Comorbidities (2 years prior), No. (%)					
Asthma	11,229 (7.3)	1,026 (6.5)	250 (6.7)	101 (3.9)	12,394 (7.2)
Bronchitis	5,820 (3.8)	884 (5.6)	205 (5.5)	115 (4.5)	6,785 (3.9)
Cardiomyopathy	4,128 (2.7)	1,374 (8.7)	319 (8.6)	222 (8.6)	5,662 (3.3)
Cancer	19,938 (12.9)	4,146 (26.2)	991 (26.6)	828 (32.1)	24,693 (14.3)
Cerebrovascular Disease	2,508 (1.6)	874 (5.5)	194 (5.2)	132 (5.1)	3,472 (2.0)
Congestive Heart Failure	8,616 (5.6)	3,668 (23.2)	909 (24.4)	657 (25.5)	12,749 (7.4)
Cirrhosis	2,583 (1.7)	849 (5.4)	228 (6.1)	163 (6.3)	3,534 (2.0)
CKD	16,723 (10.8)	4,753 (30.0)	1,192 (32.0)	967 (37.5)	22,182 (12.8)
Chronic Lung Disease	41,082 (26.5)	6,986 (44.1)	1,710 (45.9)	1,120 (43.4)	49,005 (28.4)
Cardiovascular Disease	43,506 (28.1)	9,236 (58.3)	2,197 (59.0)	1,593 (61.8)	53,997 (31.2)
Dementia	4,901 (3.2)	2,220 (14.0)	417 (11.2)	558 (21.6)	7,479 (4.3)
Diabetes	43,357 (28.0)	7,320 (46.2)	1,751 (47.0)	1,224 (47.5)	51,671 (29.9)
Comorbidity Indices, mean (SD)					
CCI within 2yrs	1.4 (1.9)	3.4 (2.9)	3.5 (2.9)	3.8 (2.9)	1.6 (2.1)
CCI, ever	2.4 (2.7)	5.1 (3.5)	5.2 (3.6)	5.6 (3.5)	2.7 (2.9)
Elixhauser within 2 yrs	0.3 (8.3)	8.5 (13.6)	9.3 (13.8)	11.0 (13.8)	1.1 (9.3)
Elixhauser, ever	1.8 (11.7)	13.2 (16.4)	13.7 (16.9)	16.5 (16.4)	2.9 (12.8)

https://doi.org/10.1371/journal.pone.0290221.t001

diagnoses of asthma and bronchitis in the 2 years prior to infection was similar among mild and non-mild infections.

Our study included detailed vaccination data (Table 1). Over 29.1% of the overall cohort were unvaccinated (neither partially or fully vaccinated). Moreover, unvaccinated Veterans accounted for a disproportionately greater percentage of deaths (41.4%) compared to fully vaccinated and recently boosted (< 5 months) Veterans, who accounted for only 14.7% of the overall cohort and 11.8% of deaths. The more advanced the patients' vaccination status, the lower their contribution to deaths (Table 1). Similar trends were observed by vaccination status for the patient groups who required hospitalization or escalation of care (Table 1).

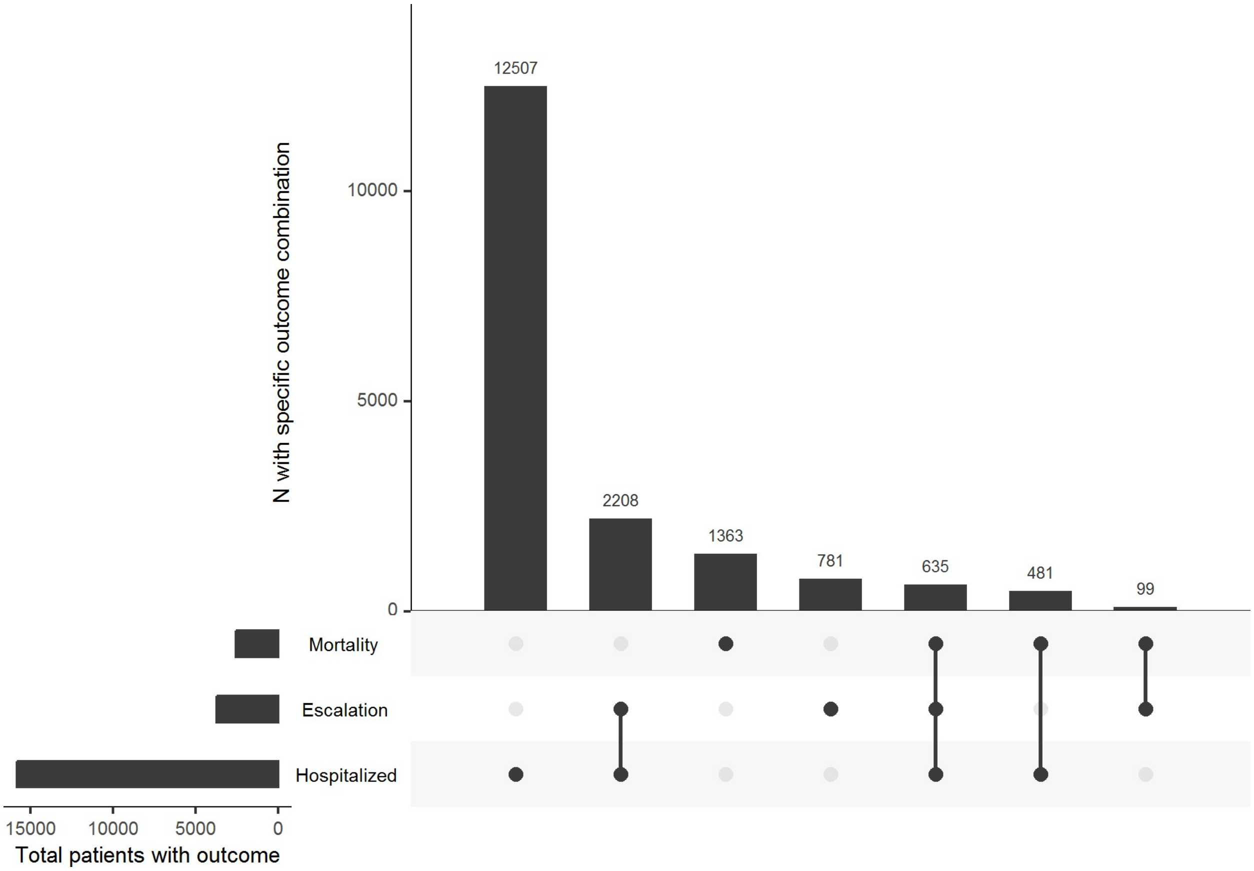

Fig 1. Upset plot of non-exclusive 30-day outcomes of interest in US Veterans. A dot in each row represents patients experiencing that outcome at any time within 30 days after testing positive. The vertical line connecting two (or more) dots represents patients who experienced two or more of the outcomes at any time within 30 days after testing positive.

https://doi.org/10.1371/journal.pone.0290221.g001

Model performance

After recursive feature selection evaluated the importance of 159 covariates, hospitalization had 25 relevant covariates, escalation of care had 75 relevant covariates, and mortality had 25 relevant covariates. The binary ML models predicted all 3 outcomes with good discrimination; all models had thresholds that maximized balance in performance, with sensitivity, specificity, and precision greater than 72% (Table 2). Consistent with its deterministic nature, death was predicted with better discrimination than the other outcomes, based on AUCs for both the receiver operator characteristic (ROC) (AUC = 0.903 95% CI [0.895, 0.911]) and normalized precision recall curves (AUC = 0.889 95% CI [0.879, 0.897]) (Fig 2). The model predicting

Table 2. Performance of machine learning models for predicting hospitalization, escalation of care, and death within 30 days after SARS-CoV-2 infection.

Outcome	Specificity [95% CI]	Sensitivity [95% CI]	Precision [95% CI]
Hospitalization	0.74 [0.73,0.74]	0.76 [0.75,0.77]	0.74 [0.74,0.75]
Escalation of care	0.72 [0.71,0.72])	0.75 [0.73,0.77]	0.73 [0.72,0.73]
Mortality	0.83 [0.83,0.83]	0.82 [0.79,0.85]	0.83 [0.82,0.83]

https://doi.org/10.1371/journal.pone.0290221.t002

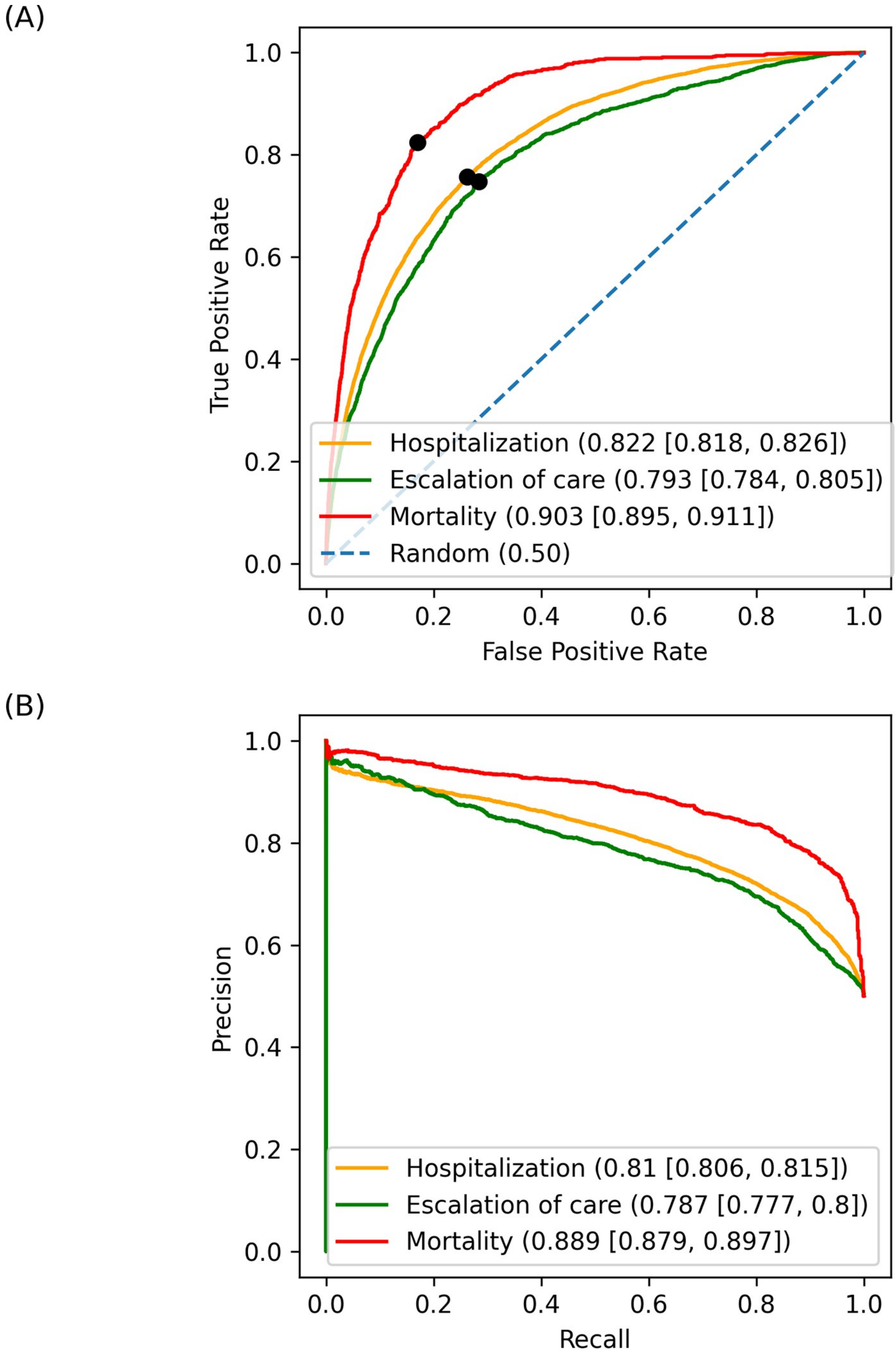

Fig 2. Classification performance curves with respective area under curve (AUC) and 95% confidence intervals. (A) Receiver Operating Characteristic (ROC) curve for each model with respective false positive and true positive rates at the classification thresholds indicated by black dots. (B) Normalized precision recall curve for each 30-day outcome.

https://doi.org/10.1371/journal.pone.0290221.g002

hospitalization had better discrimination than the model for the need for escalation of care (hospitalization: AUC = 0.822 95% CI [0.818, 0.826]; escalated hospital care: AUC = 0.793 95% CI [0.784, 0.805]) (Fig 2).

Model interpretation

We evaluated the covariates that most predicted risks of hospitalization, escalation of care, and mortality within 30 days of a SARS-CoV-2 positive test during the observation period. Feature importance was measured as the fraction of total error reduction for a given covariate (Fig 3). We generated SHAP summary plots to show the impact of covariate values on predictive output (S1 Fig). Advanced age was the second most predictive covariate for hospitalization (Fig 3A and S1A Fig). It was also the most predictive covariate for escalation of care (Fig 3B and S1B Fig) and mortality, accounting for more than 50% of relative importance (Fig 3C and S1C Fig).

Weighted indices of comorbid illnesses, the Charlson Comorbidity index (CCI) [14] and Elixhauser index [15], were more robust predictors of the adverse outcomes than individual cardiometabolic, renal, and respiratory conditions (Fig 3). BMI was highly predictive of the outcomes; BMI was inversely proportional to predicted risk, based upon SHAP analysis (Fig 3 and S1 Fig). Veterans taking an oral anticoagulant at any time in the two years prior to testing positive for SARS-CoV-2 had higher risks of hospitalization and need for escalation of care (Fig 3A, 3B and S1A, S1B Fig). Patients who had been prescribed vasopressors at any time in the prior two years had a higher predicted risk for escalation of care, while patients on the diuretic, furosemide, had higher predicted risk for mortality (Fig 3B, 3C and S1B, S1C Fig).

Fully vaccinated and boosted patients had lower predicted risks of hospitalization, escalation of care, and death at 30 days. Additionally, unknown blood type and alternative insurance were among the most significant predictors of a lower risk for hospitalization, while residing in non-rural areas and being male were among the most important predictors of mortality risk (Fig 3A, 3C and S1A, S1C Fig).

Projected impact of risk-prioritized vaccination strategies

To project the impact of targeted vaccination on adverse outcomes using the prediction models, we examined the unvaccinated subset (n = 22,082) from the test cohort (n = 92,080). We projected the number of adverse outcomes for three *in silica* scenarios: (1) vaccination of all Veterans within the unvaccinated subset, (2) random vaccination of 20% of the unvaccinated Veterans, and (3) vaccination of only the Veterans in the top quintile of predicted risk for adverse outcomes (Table 3). Using sensitivity tradeoff curves (S2 Fig), we observed a step-up of predicted risk at the top quintile. Therefore, we selected the cut-off to be the top quintile of the population. In turn, our modeling projected the optimum impact of risk-prioritized vaccination strategy. Full vaccination of the entire unvaccinated population in our test set was predicted to reduce hospitalizations by 82.1% (from 1,698 to 304), escalations of care by 82.9% (from 351 to 60), and deaths by 84.4% (from 179 to 28.1). When a random 20% of the unvaccinated population was vaccinated in the projection modeling, hospitalizations were reduced from 1,698 to 1,504 (11.4% reduction), escalations of care from 351 to 313 (10.8%), and deaths from 179 to 161 (10.1%). When vaccinating the patients in the top quintile (20%) of the highest risk for adverse outcomes, hospitalizations were reduced from 1,698 to 1,017 (40.1%), escalations of care from 351 to 233 (33.6%), and deaths from 179 to 101 (43.6%).

Discussion

In a national cohort of 172,814 US Veterans who tested positive for SARS-CoV-2 during the Omicron surge, we demonstrated the most robust prediction discrimination to date for 30-day risk for hospitalization, escalation of care, and mortality after COVID-19 infection, using ML methods. Our ML models leveraged data including detailed vaccination status during the Omicron surge. We identified predictors for, and projected subgroups of, high-risk individuals

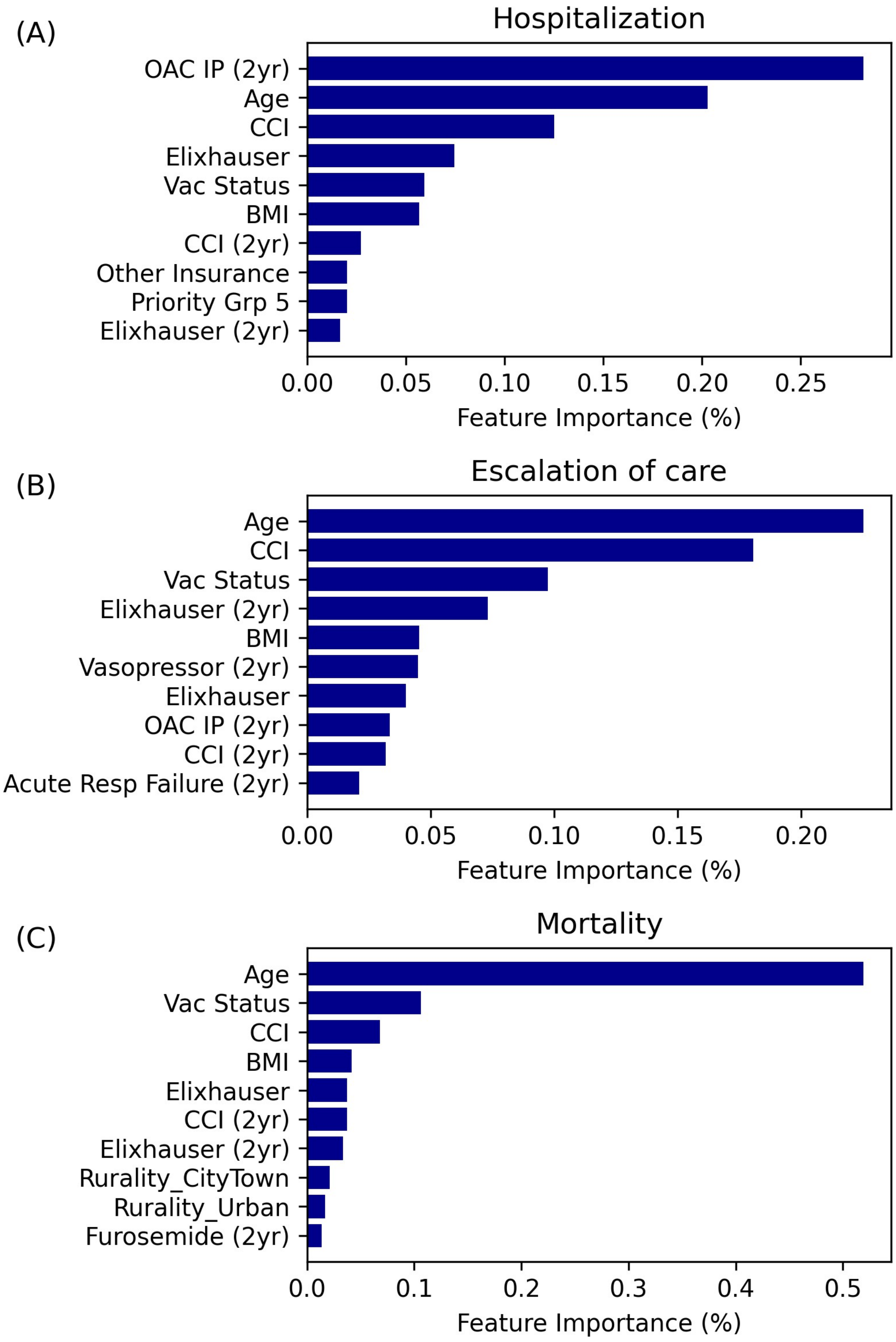

Fig 3. Clinical feature importance plot. (A) hospitalization, (B) escalation of care, and (C) mortality. Feature importance values for each of the three outcomes of interest are presented as a percentage, which is indicative of the fraction of error reduction that a given feature contributed to the model.

https://doi.org/10.1371/journal.pone.0290221.g003

Table 3. Observations and projections for occurrences for hospitalization, escalation of care, and mortality, for three vaccination scenarios.

	Observed	Projections (boostrap = 100)		
Outcome (30-day Risk)	Unvaccinated (n = 22,082)	Vaccination of All Unvaccinated [95% CI]	Vaccination of Random 20% [95% CI]	Vaccination in Top Quintile (20th %ile) Risk [95% CI]
Hospitalization	1,698	304.05 [303.39, 304.71]	1,504.19 [1503.63, 1504.75]	1,017.14 [1016.64, 1017.63]
Escalation of Care	351	59.56 [59.27, 59.85]	313.03 [312.77, 313.28]	233.13 [232.9, 233.36]
Mortality	179	28.07 [27.87, 28.27]	161.13 [160.96, 161.3]	101.36 [101.17, 101.54]

https://doi.org/10.1371/journal.pone.0290221.t003

who stand to benefit the most from advancing vaccination status. Prioritizing vaccination of individuals in the highest quintile of predicted risk for hospitalization or death was projected to produce greater than 3.5-fold projected reductions in hospitalization and death, compared to randomly vaccinating 20% of the population.

Previous prediction models, including those developed in the VHA, utilized data collected prior to the emergence of the Omicron SARS-CoV-2 variant [9–12]. A large retrospective analysis of over 1.5 million vaccinated patients in the VHA showed relatively low rates of breakthrough infections and related complications such as pneumonia and death [21]. This statistically powerful investigation excluded unvaccinated individuals and anyone with a prior history of COVID-19 infection, and risk prediction modeling was not a primary focus of that report. Although a prior smaller study incorporated vaccination into ML risk prediction modeling for COVID-19 [22], our study incorporated stratified vaccination status, which reflects degree of protection through number and timing of primary and booster vaccines in an ML-driven risk prediction model.

Compared to recent studies, ML models in the present study demonstrated more robust discrimination by AUC in predicting 30-day risk for hospitalization (AUC 0.822), escalation of care (AUC 0.793), and mortality (AUC 0.903) with COVID-19 infection. Two prior studies derived from cohorts of ~4,500 patients each demonstrated lower AUCs (0.804 and 0.813) for predicting hospitalization [23, 24]. A previous model developed from a large VHA cohort of 7,635,064 (both infected and non-infected) with an observation window from May 21 to November 2, 2020 predicted 30-day mortality with a validation AUC of 0.836 (95% CI, 82.0%-85.3%) [9]. In addition, a recent study of 1,201 patients who contracted SARS-CoV-2 in Spain in 2020 predicted 30-day mortality with an AUC of 0.872 [25]. Commonly identified covariates in prior studies, advanced age and higher medical co-morbidity indices, were associated with higher risks for the adverse outcomes of interest in our models [9–11]. Our models identified a general inverse association between BMI and predicted risk for adverse outcomes. This contrasts a prior meta-analysis that demonstrated that higher BMI (and visceral adiposity) correlates with a higher risk of hospitalization, mortality, and other adverse outcomes such as admission to ICU and need for mechanical ventilation [26].

Consistent with prior vaccine trials [27], our study indicated that vaccination reduces hospitalizations, escalation of care, and deaths. Individuals who were fully vaccinated and boosted within 5 months from testing SARS-CoV-2 positive had the greatest projected protection. Use of oral anticoagulants in the two years prior to current infection strongly predicted 30-day hospitalization and escalation of care. The biological basis of this observation may be related to the underlying medical conditions that warranted anticoagulation or to specific effects of the anticoagulants themselves. Notably, baseline furosemide use was associated with a higher risk of death, suggesting that underlying heart failure or volume-expanded states are important determinants of infection severity in Omicron infections.

Limitations

The present findings in this national study of US Veterans may not be broadly applicable to the general population. Consistent with the US Veteran population, our study cohort was predominantly male and white with greater medical comorbidity and lower socioeconomic status than the general US population. The relevance of the models remains limited for racial/ethnic minority communities who have borne a disproportionate burden during the pandemic. However, the methodology used here can be applied and adapted to other populations or health care systems. Additionally, while some recent work has sought to remove confounding effects from machine learning models in imaging [28, 29], these statistical approaches can lead to biases in estimating predictive modeling performances [30, 31]. While statistical analysis is best suited for estimating the causality of features on outcome, here we sought to optimize robust predictive performance through machine learning and highlight predictive features. For vaccine projections, all outcomes of interest were assumed to be the result of SARS-CoV-2 infection. While the VHA COVID-19 Shared Data Resource database captures all deaths, it does not capture hospitalizations and care received outside the VA. This may explain why having other non-VHA insurance was associated with lower rates of 30-day hospitalization given that patients with non-VHA insurance may have sought care outside the VA. The VHA COVID-19 Shared Data Resource database also does not establish whether SARS-CoV-2/COVID-19 is the reason for hospitalization, escalation of care, or death. Determining this is challenging. Our modeling also does not include laboratory or imaging data; these data have been shown to have robust predictive value post index date [32–35]. Finally, the model results were most relevant to Omicron variants and sub-variants and may not be relevant to other pathogenetic SARS-CoV-2 variants.

Conclusions

Our ML risk prediction modeling approach provides robust discrimination in predicting hospitalization, escalated hospital care and death within 30 days of testing positive for SARS-CoV-2 infection during a recent observation period in which Omicron variants are the major cause of COVID-19. It can inform health care system vaccination and resource allocation decisions by characterizing individuals and subpopulations at low-to-high risk for 30-day hospitalization, escalated hospital care or death, and identifying those who might benefit least-to-most from preventive intervention. While this modeling was developed specifically for the Omicron variant surge, analogous modeling can be developed and implementable rapidly in real-time to guide vaccination strategies and resource allocation during future COVID-19 surges.

Supporting information

S1 Table. Covariates used in predictive modeling. A table of all potential covariates that were investigated with a brief definition.
(XLSX)

S2 Table. Comparison of modeling approaches. A table of the performance metrics given here as ROCAUC for different machine learning algorithms for each severity outcome.
(XLSX)

S1 Fig. SHAP summary plots for 30-day outcomes of interest. (A) hospitalization, (B) escalation of care, and (C) mortality. Covariates are listed in order of highest to lowest impact (based on absolute mean SHAP value) along the y-axis. Each blue or red point represents a patient's specified covariate value; that value is color coded in a heat map fashion per the

legend. The x-axis is the SHAP value for the specific covariate, with SHAP values greater than 0 indicating higher predicted risk contribution and values less than 0 indicating lower predicted risk contribution for the given outcome.
(TIF)

S2 Fig. Sensitivity tradeoff curve. A plot of the percentile of a target outcome as it relates to the risk percentile for our test population.
(TIF)

Acknowledgments

The authors would like to thank Hui Wang, Laurel Stell, and Wu Fan, for their contributions to this work.

Author Contributions

Conceptualization: Steven Cogill, Shriram Nallamshetty, Mei-Chiung Shih, Jennifer S. Lee.

Data curation: Steven Cogill, Kent Heberer.

Formal analysis: Steven Cogill, Kent Heberer.

Funding acquisition: Julie Lynch, Kyung Min Lee, Mihaela Aslan, Mei-Chiung Shih, Jennifer S. Lee.

Investigation: Steven Cogill, Jennifer S. Lee.

Methodology: Steven Cogill, Shriram Nallamshetty, Kent Heberer, Mei-Chiung Shih, Jennifer S. Lee.

Project administration: Shriram Nallamshetty, Natalie Fullenkamp, Mihaela Aslan, Mei-Chiung Shih, Jennifer S. Lee.

Resources: Steven Cogill, Julie Lynch, Kyung Min Lee, Mihaela Aslan, Mei-Chiung Shih, Jennifer S. Lee.

Software: Steven Cogill, Kent Heberer.

Supervision: Shriram Nallamshetty, Natalie Fullenkamp, Julie Lynch, Mihaela Aslan, Mei-Chiung Shih, Jennifer S. Lee.

Validation: Steven Cogill, Jennifer S. Lee.

Visualization: Steven Cogill, Shriram Nallamshetty, Kent Heberer, Mihaela Aslan, Mei-Chiung Shih.

Writing – original draft: Steven Cogill, Shriram Nallamshetty.

Writing – review & editing: Steven Cogill, Shriram Nallamshetty, Natalie Fullenkamp, Kent Heberer, Julie Lynch, Kyung Min Lee, Mihaela Aslan, Mei-Chiung Shih, Jennifer S. Lee.

References

1. WHO Coronavirus (COVID-19) Dashboard. [cited 2 Jun 2022]. Available: https://covid19.who.int
2. CDC COVID Data Tracker. [cited 18 Mar 2022]. Available: https://covid.cdc.gov/covid-data-tracker/#variant-proportions
3. Iuliano AD, Brunkard JM, Boehmer TK, Peterson E, Adjei S, Binder AM, et al. Trends in Disease Severity and Health Care Utilization During the Early Omicron Variant Period Compared with Previous SARS-CoV-2 High Transmission Periods—United States, December 2020-January 2022. MMWR Morb Mortal Wkly Rep. 2022; 71: 146–152. https://doi.org/10.15585/mmwr.mm7104e4 PMID: 35085225

4. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. [cited 8 Jun 2022]. Available: https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern

5. VanBlargan LA, Errico JM, Halfmann PJ, Zost SJ, Crowe JE, Purcell LA, et al. An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. Nat Med. 2022; 28: 490–495. https://doi.org/10.1038/s41591-021-01678-y PMID: 35046573

6. Abdullah F, Myers J, Basu D, Tintinger G, Ueckermann V, Mathebula M, et al. Decreased severity of disease during the first global omicron variant covid-19 outbreak in a large hospital in tshwane, south africa. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2022; 116: 38–42. https://doi.org/10.1016/j.ijid.2021.12.357 PMID: 34971823

7. Chen J, Wang R, Gilby NB, Wei G-W. Omicron Variant (B.1.1.529): Infectivity, Vaccine Breakthrough, and Antibody Resistance. J Chem Inf Model. 2022; 62: 412–422. https://doi.org/10.1021/acs.jcim.1c01451 PMID: 34989238

8. Nyberg T, Ferguson NM, Nash SG, Webster HH, Flaxman S, Andrews N, et al. Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: a cohort study. The Lancet. 2022; 399: 1303–1312. https://doi.org/10.1016/S0140-6736(22)00462-7 PMID: 35305296

9. Ioannou GN, Green P, Fan VS, Dominitz JA, O'Hare AM, Backus LI, et al. Development of COVIDVax Model to Estimate the Risk of SARS-CoV-2–Related Death Among 7.6 Million US Veterans for Use in Vaccination Prioritization. JAMA Netw Open. 2021; 4: e214347. https://doi.org/10.1001/jamanetworkopen.2021.4347 PMID: 33822066

10. Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, et al. Prediction for Progression Risk in Patients With COVID-19 Pneumonia: The CALL Score. Clin Infect Dis Off Publ Infect Dis Soc Am. 2020; 71: 1393–1399. https://doi.org/10.1093/cid/ciaa414 PMID: 32271369

11. Jung C, Excoffier J-B, Raphaël-Rousseau M, Salaün-Penquer N, Ortala M, Chouaid C. Evolution of hospitalized patient characteristics through the first three COVID-19 waves in Paris area using machine learning analysis. PloS One. 2022; 17: e0263266. https://doi.org/10.1371/journal.pone.0263266 PMID: 35192649

12. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients With COVID-19. JAMA Intern Med. 2020; 180: 1081–1089. https://doi.org/10.1001/jamainternmed.2020.2033 PMID: 32396163

13. UpSetR: an R package for the visualization of intersecting sets and their properties | Bioinformatics | Oxford Academic. [cited 14 Jul 2022]. Available: https://academic.oup.com/bioinformatics/article/33/18/2938/3884387

14. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987; 40: 373–383. https://doi.org/10.1016/0021-9681(87)90171-8 PMID: 3558716

15. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity Measures for Use with Administrative Data. Med Care. 1998; 36: 8–27. https://doi.org/10.1097/00005650-199801000-00004 PMID: 9431328

16. LeDell E, Poirier S. H2O AutoML: Scalable Automatic Machine Learning. 7th ICML Workshop Autom Mach Learn AutoML. 2020. Available: https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf

17. mateuszbuda. Machine Learning Statistical Utils. 2022. Available: https://github.com/mateuszbuda/ml-stat-util

18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011; 12: 2825–2830.

19. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. arXiv; 2017 Nov. Report No.: arXiv:1705.07874. https://doi.org/10.48550/arXiv.1705.07874

20. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Covid-19 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. N Engl J Med. 2022; 386: 1532–1546. https://doi.org/10.1056/NEJMoa2119451 PMID: 35249272

21. Kelly JD, Leonard S, Hoggatt KJ, Boscardin WJ, Lum EN, Moss-Vazquez TA, et al. Incidence of Severe COVID-19 Illness Following Vaccination and Booster With BNT162b2, mRNA-1273, and Ad26.COV2.S Vaccines. JAMA. 2022; 328: 1427–1437. https://doi.org/10.1001/jama.2022.17985 PMID: 36156706

22. Ong SWX, Tham SM, Koh LP, Dugan C, Khoo BY, Ren D, et al. External validation of the PRIORITY model in predicting COVID-19 critical illness in vaccinated and unvaccinated patients. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2022; 28: 884.e1–884.e3. https://doi.org/10.1016/j.cmi.2022.01.031 PMID: 35150879

23. Jehi L, Ji X, Milinovich A, Erzurum S, Merlino A, Gordon S, et al. Development and validation of a model for individualized prediction of hospitalization risk in 4,536 patients with COVID-19. PLOS ONE. 2020; 15: e0237419. https://doi.org/10.1371/journal.pone.0237419 PMID: 32780765

24. Willette AA, Willette SA, Wang Q, Pappas C, Klinedinst BS, Le S, et al. Using machine learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK Biobank cohort study. Sci Rep. 2022; 12: 7736. https://doi.org/10.1038/s41598-022-07307-z PMID: 35545624

25. Reina Reina A, Barrera JM, Valdivieso B, Gas M-E, Maté A, Trujillo JC. Machine learning model from a Spanish cohort for prediction of SARS-COV-2 mortality risk and critical patients. Sci Rep. 2022; 12: 5723. https://doi.org/10.1038/s41598-022-09613-y PMID: 35388055

26. Demeulemeester F, de Punder K, van Heijningen M, van Doesburg F. Obesity as a Risk Factor for Severe COVID-19 and Complications: A Review. Cells. 2021; 10: 933. https://doi.org/10.3390/cells10040933 PMID: 33920604

27. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. N Engl J Med. 2020; 383: 2603–2615. https://doi.org/10.1056/NEJMoa2034577 PMID: 33301246

28. Snoek L, Miletić S, Scholte HS. How to control for confounds in decoding analyses of neuroimaging data. NeuroImage. 2019; 184: 741–760. https://doi.org/10.1016/j.neuroimage.2018.09.074 PMID: 30268846

29. Zhao Q, Adeli E, Pohl KM. Training confounder-free deep learning models for medical applications. Nat Commun. 2020; 11: 6010. https://doi.org/10.1038/s41467-020-19784-9 PMID: 33243992

30. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci. 2009; 12: 535–540. https://doi.org/10.1038/nn.2303 PMID: 19396166

31. Jamalabadi H, Alizadeh S, Schönauer M, Leibold C, Gais S. Classification based hypothesis testing in neuroscience: Below-chance level classification rates and overlooked statistical properties of linear parametric classifiers. Hum Brain Mapp. 2016; 37: 1842–1855. https://doi.org/10.1002/hbm.23140 PMID: 27015748

32. Alle S, Kanakan A, Siddiqui S, Garg A, Karthikeyan A, Mehta P, et al. COVID-19 Risk Stratification and Mortality Prediction in Hospitalized Indian Patients: Harnessing clinical data for public health benefits. PLoS ONE. 2022; 17: e0264785. https://doi.org/10.1371/journal.pone.0264785 PMID: 35298502

33. Butler L, Karabayir I, Samie Tootooni M, Afshar M, Goldberg A, Akbilgic O. Image and structured data analysis for prognostication of health outcomes in patients presenting to the ED during the COVID-19 pandemic. Int J Med Inf. 2021; 158: 104662. https://doi.org/10.1016/j.ijmedinf.2021.104662 PMID: 34923448

34. Ortiz A, Trivedi A, Desbiens J, Blazes M, Robinson C, Gupta S, et al. Effective deep learning approaches for predicting COVID-19 outcomes from chest computed tomography volumes. Sci Rep. 2022; 12: 1716. https://doi.org/10.1038/s41598-022-05532-0 PMID: 35110593

35. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomography. Cell. 2020; 181: 1423–1433.e11. https://doi.org/10.1016/j.cell.2020.04.045 PMID: 32416069