

20 ⁵VA Clinical Epidemiology Research Center (CERC), VA Connecticut Healthcare System, West
21 Haven, CT

22 ⁶Department of Medicine, Yale University School of Medicine, New Haven, CT

23 ⁷Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA

24 ⁸Department of Medicine, Division of Endocrinology, Gerontology, and Metabolism, and by
25 courtesy, of Epidemiology and Population Health, Stanford University School of Medicine,
26 Stanford, CA

27
28 *Corresponding author
29 Jennifer S. Lee, MD, PhD (corresponding author)
30 Associate Chief of Staff for Research, VA Palo Alto Health Care System
31 Associate Professor of Medicine, Stanford School of Medicine
32 Email: Jennifer.Lee23@va.gov
33 3801 Miranda Avenue
34 Palo Alto, CA 94304

35
36 Number of Figures and Tables: 3 Figures, 3 Tables, 1 Supplementary Table, 1 Supplementary
37 Figure

38 Number of References: 29

39

40

41

42

43

44

45 **Abstract**

46 The Omicron SARS-CoV-2 variant continues to strain healthcare systems. Developing
47 tools that facilitate the identification of patients at highest risk of adverse outcomes is a priority.
48 The study objectives are to develop population-scale predictive models that: 1) identify predictors
49 of adverse outcomes with Omicron surge SARS-CoV-2 infections, and 2) predict the impact of
50 prioritized vaccination of high-risk groups for said outcome. We prepared a retrospective
51 longitudinal observational study of a national cohort of 192,984 patients in the U.S. Veteran Health
52 Administration who tested positive for SARS-CoV-2 from January 15 to August 15, 2022. We
53 utilized sociodemographic characteristics, comorbidities, vaccination status, and prior COVID-19
54 infections, at time of testing positive for SARS-CoV-2 to predict hospitalization, escalation of care
55 (high-flow oxygen, mechanical ventilation, vasopressor use, dialysis, or extracorporeal membrane
56 oxygenation), and death within 30 days. Machine learning models demonstrated that advanced
57 age, high comorbidity burden, lower body mass index, unvaccinated status, prior SARS-CoV-2
58 infection, and oral anticoagulant use were the important predictors of hospitalization and escalation
59 of care. Similar factors predicted death. However, prior SARS-CoV-2 infection was associated
60 with lower 30-day mortality, and anticoagulant use did not predict mortality risk. The all-cause
61 death model showed the highest discrimination (Area Under the Curve (AUC) = 0.895, 95%
62 Confidence Interval (CI): 0.885, 0.906) followed by hospitalization (AUC = 0.829, CI: 0.825,
63 0.834), then escalation of care (AUC=0.805, CI: 0.795, 0.814). Assuming a vaccine efficacy range
64 of 70.8 to 78.7%, our simulations projected that targeted prevention in the highest risk group may
65 have reduced 30-day hospitalization, care escalation, and death in more than 2 of 5 unvaccinated
66 patients.

67 **Introduction**

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

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

89 **Materials and methods**

90 **Study cohort**

91 Our study cohort consisted of all 192,984 Veterans who tested positive for COVID-19
92 between January 15 and August 15, 2022, as captured by the VHA's COVID-19 Shared Data
93 Resource with data curation within the VHA's Corporate Data Warehouse (CDW). Data were
94 accessed on 10/4/22 for research purposes. No new data were collected and no direct patient (or
95 participant) contact took place. Patients' curated electronic health records in the VHA's CDW
96 were analyzed behind the VHA secured firewall as part of the VHA research data initiative,
97 Leveraging Electronic Health Information to Advance Precision medicine (LEAP, CSP#2012),
98 which has been approved by VHA's Central Institutional Review Board and Research &
99 Development Committees at 3 VA Medical Centers (Salt Lake City, Palo Alto, and West
100 Haven). The date of the positive test is defined as the index date. For the selected cohort within
101 the data resource, there were no missing data for the selected fields and unknown covariates were
102 indicated as such. Patients outside the age range of 18 to 100 or outside the Body Mass Index
103 (BMI) range of 15 to 100 were excluded from the analysis.

104 **Study outcomes**

105 We predicted the risk of developing one of the following three distinct, non-mutually
106 exclusive clinical outcomes representing SARS-CoV-2 severity within 30 days of infection: (i)
107 hospitalization, (ii) escalation of care (defined as the need for high-flow supplemental oxygen,
108 mechanical ventilation, vasopressors, renal replacement therapy [with no prior dialysis in the
109 preceding two years], or extracorporeal membrane oxygenation [ECMO]), and (iii) all-cause

110 mortality. Patients who tested positive for SARS-CoV-2 were deemed to have ‘mild’ infection if
111 they did not experience any of the three outcomes of interest within 30 days of infection. The Upset
112 plot was generated using the UpsetR package [13].

113 **Clinical Features**

114 A total of 159 patient characteristics including medical comorbidities, demographic data,
115 vaccination status, prior COVID-19 infection status, and comorbidity indices were available for
116 each patient prior to feature selection. The medical history included pre-existing conditions,
117 procedures, and medications. All medical history values were classified using a Boolean system
118 for presence or absence of the specific medical condition within two years prior to the current
119 COVID-19 infection. Demographic and clinical data employed in the modeling included age, sex,
120 race/ethnicity, blood type, BMI, veteran status, whether overweight at index date, rurality of
121 current residence, and veteran priority status (a surrogate for income status and benefits eligibility).
122 These covariates were multimodal (float, categorical and Boolean). Vaccination status was
123 represented as a categorical score from 0 to 5 as follows: 0=no vaccination, 1=partial-mRNA
124 vaccination, 2=full vaccination (two doses of mRNA or a single dose of viral vector-based vaccine)
125 > 5 months from index date, 3=fully-vaccinated and boosted >5 months prior to the index date,
126 4=fully-vaccinated <5 months prior to the index date, 5=fully-vaccinated and boosted <5 months
127 prior to the index date. Vaccines given outside of the VHA were available in the VHA COVID-19
128 Shared Data Resource and reflected in our dataset. Vaccination status accounted for a two-week
129 efficacy window. A veteran was considered to be positive for prior COVID-19 infection if the
130 prior positive test was more than two weeks before the index date of the current infection. Medical
131 comorbidity burden was assessed by Charlson Comorbidity Index (CCI) [14] and Elixhauser Index
132 [15] scores for the two years prior to infection. An overall CCI and Elixhauser index score was
133 also determined. A complete list of covariates is included in S1 Table.

134 **Model Development and Performance**

135 For each of the 3 main outcomes of interest, we developed a distinct binary model that
136 incorporated 159 unique covariate features using gradient boosting automated machine learning
137 methods. A recursive feature elimination approach was used to find the most parsimonious models.
138 Our data was split chronologically with training/validation data from January 15, 2022 to April
139 15, 2022 and our test data from April 16, 2022 to August 15, 2022. Covariates with variance lower
140 than 1% within the training set were removed, and non-binary values were scaled from 0 to 1.

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

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

157

158 **Model Interpretation and Applications**

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

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

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

179 **Results**

180 **Patient population and clinical predictors of COVID-19**
 181 **infection severity**

182 In a national VHA cohort of 192,984 Veterans who tested positive for SARS-CoV-2
 183 during a period in which the Omicron variant predominated (January 15-August 15, 2022), the
 184 median age was 62 years and 83.8% were men (Table 1). The racial/ethnic composition of the
 185 cohort was typical for a US Veteran population; 65.2% of the patients were white, 19.5% were
 186 black, and 9% were Hispanic. Asian, Native Hawaiian or Pacific Islander, and American Indian
 187 or Alaskan Native Veterans each represented approximately 1 % of the cohort. (Table 1).

188

189 **Table 1. 30-day outcomes after a positive SARS-CoV-2 test.**

	Mild	Hospitalized	Escalation	Death	Overall
Characteristics	n=170,422 (88.3%)	n=20,267 (10.5%)	n=4,756 (2.5%)	n=2,635 (1.4%)	n=192,984
Age, median [IQR]	60 [46, 72]	72 [63, 78]	72 [64, 77]	77 [72, 85]	62 [48, 73]
BMI, mean (SD)	30.3 (6.2)	28.3 (6.9)	28.6 (7.1)	26.7 (6.7)	30.1 (6.3)
Men, No. (%)	140,223 (82.3)	19,149 (94.5)	4,515 (94.7)	2,583 (98.0)	161,554 (83.7)
<u>Race, No. (%)</u>					
White	110,204 (64.7)	13,892 (68.5)	3,406 (71.4)	2,050 (77.8)	125,858 (65.2)
Black	32,741 (19.2)	4,525 (22.3)	901 (18.9)	311 (11.8)	37,544 (19.5)
Asian	2,696 (1.6)	128 (0.6)	36 (0.8)	15 (0.6)	2,844 (1.5)
Native American/Alaska Native	1,388 (0.8)	169 (0.8)	39 (0.8)	22 (0.8)	1,576 (0.8)
Native Hawaiian/Other Pacific Islander	1,833 (1.1)	146 (0.7)	43 (0.9)	23 (0.9)	2,005 (1.0)
Unknown	21,560 (12.7)	1,407 (6.9)	344 (7.2)	214 (8.1)	23,157 (12.0)
Hispanic or Latino, No. (%)	16,041 (9.4)	1,687 (8.3)	420 (8.8)	177 (6.7)	17,864 (9.3)

Previous Covid-19 Infection, No. (%)	16,247 (9.5)	4,352 (21.5)	1,173 (24.6)	104 (3.9)	20,721 (10.7)
<u>Vaccination status, No. (%)</u>					
0-Unvaccinated	54,027 (31.7)	5,987 (29.5)	1,599 (33.5)	1,099 (41.7)	60,900 (31.6)
1-Partial mRNA (1 dose)	3,243 (1.9)	455 (2.2)	96 (2.0)	65 (2.5)	3,745 (1.9)
2-Fully Vaccinated, > 5 months prior	44,417 (26.1)	5,356 (26.4)	1,246 (26.1)	741 (28.1)	50,404 (26.1)
3-Fully Vaccinated, with Booster > 5 months prior	41,886 (24.6)	5,118 (25.3)	1,049 (22.0)	374 (14.2)	47,399 (24.6)
4-Fully Vaccinated, <5 months prior	3,317 (1.9)	379 (1.9)	87 (1.8)	39 (1.5)	3,735 (1.9)
5-Fully Vaccinated with Booster <5 months prior	23,532 (13.8)	2,972 (14.7)	692 (14.5)	317 (12.0)	26,801 (13.9)
<u>Comorbidities (2 years prior), No. (%)</u>					
Asthma	12,535 (7.4)	1,394 (6.9)	337 (7.1)	102 (3.9)	14,074 (7.3)
Bronchitis	7,059 (4.1)	1,270 (6.3)	306 (6.4)	116 (4.4)	8,417 (4.4)
Cardiomyopathy	4,529 (2.7)	1,782 (8.8)	432 (9.1)	221 (8.4)	6,475 (3.4)
Cancer	21,769 (12.8)	5,346 (26.4)	1,298 (27.2)	842 (32.0)	27,729 (14.4)
Cerebrovascular Disease	2,725 (1.6)	1,124 (5.5)	254 (5.3)	134 (5.1)	3,938 (2.0)
Congestive Heart Failure	9,467 (5.6)	4,747 (23.4)	1,205 (25.3)	672 (25.5)	14,684 (7.6)
Cirrhosis	2,881 (1.7)	1,115 (5.5)	296 (6.2)	171 (6.5)	4,103 (2.1)
CKD	18,242 (10.7)	6,183 (30.5)	1,582 (33.2)	982 (37.3)	25,139 (13.0)
Chronic Lung Disease	45,307 (26.6)	8,996 (44.4)	2,220 (46.6)	1,128 (42.8)	55,255 (28.6)
Cardiovascular Disease	47,562 (27.9)	11,925 (58.8)	2,844 (59.6)	1,627 (61.7)	60,766 (31.5)
Dementia	5,360 (3.1)	2,863 (14.1)	521 (10.9)	552 (20.9)	8,571 (4.4)
Diabetes	47,385 (27.8)	9,494 (46.8)	2,292 (48.1)	1,247 (47.3)	57,896 (30.0)
<u>Comorbidity Indices, mean (SD)</u>					
CCI within 2 yrs	1.4 (1.9)	3.5 (2.9)	3.6 (3.0)	3.8 (2.9)	1.6 (2.2)
CCI, ever	2.4 (2.7)	5.2 (3.5)	5.3 (3.6)	5.6 (3.5)	2.7 (3.0)
Elixhauser within 2 yrs	4.8 (8.2)	15.1 (14.5)	15.9 (15.0)	16.0 (15.1)	6.0 (9.7)
Elixhauser, ever	10.2 (12.7)	24.7 (17.6)	25.2 (18.3)	26.1 (17.9)	11.8 (2.2)

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

191

192 Overall, 88.3% of Veterans had mild SARS-CoV-2 infection. Among Veterans who tested
193 positive for SARS-CoV-2, 10.5% required hospitalization, 2.5% needed escalation of care, and
194 1.4% died (Table 1 and Fig 1). In the subset of hospitalized infected patients, a higher percentage
195 required escalation of care (15.3%) and died (6.0%) compared to the overall cohort (Fig 1).
196 Patients who died or required hospitalization and/or escalation of care were older and more likely
197 to be male. Conversely, patients who had mild infections had a higher body mass index (BMI)
198 than those who did not (Table 1). A higher percentage of patients who died were white, compared
199 to the overall cohort (77.8% vs 65.2%). In contrast, a lower percentage of patients who died were
200 black, compared to those in the overall cohort (11.8% vs. 19.5%) (Table 1).

201

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

206

207 Patients with non-mild infections had significantly higher prevalence of diabetes,
208 congestive heart failure, cerebrovascular disease, chronic kidney disease, and cirrhosis. Dementia
209 was also more prevalent among patients who required hospitalization, required escalation of care,
210 or died within 30 days after testing positive. While chronic lung disease also was more prevalent,
211 diagnoses of asthma and bronchitis in the 2 years prior to infection was not significantly different
212 among those with any of the three outcomes of interest. The database used for this study also
213 included information on prior SARS-CoV-2 infection as well as vaccination status (Table 1).

214 Approximately 10.7% of the overall cohort had a history of SARS-CoV-2 infection. A higher
215 percentage of patients with prior SARS-CoV-2 infection required hospitalization (21.5% vs 10.5%
216 for overall cohort) or escalation of care (24.6% vs 2.5% for overall cohort). In contrast, a lower
217 percentage (3.9%) of patients with a history of SARS-CoV-2 died within 30 days after the current
218 SAR-CoV-2 infection.

219 Our study also included detailed vaccination data (Table 1). Over 31.6% of the overall
220 cohort were unvaccinated (neither partially or fully vaccinated). Moreover, unvaccinated Veterans
221 accounted for a disproportionately greater percentage of deaths (41.7%) compared to fully
222 vaccinated and recently boosted (< 5 months) Veterans, who accounted for only 13% of the overall
223 cohort and 12% of deaths. The more advanced the patients' vaccination status, the lower their
224 contribution to deaths (Table 1). Similar trends were observed by vaccination status for the patient
225 groups who required hospitalization or escalation of care (Table 1).

226 **Model performance**

227 After recursive feature selection evaluated the importance of 159 covariates,
228 hospitalization had 20 relevant covariates, escalation of care had 25 relevant covariates, and
229 mortality had 15 relevant covariates. The binary ML models predicted all 3 outcomes with good
230 discrimination; all models had thresholds that maximized balance in performance, with sensitivity,
231 specificity, and precision greater than 73% (Table 2). Consistent with its deterministic nature,
232 death was predicted with better discrimination than the other outcomes, based on AUCs for both
233 the receiver operator characteristic (ROC) (AUC = 0.895 95% CI [0.885, 0.906]) and normalized
234 precision recall curves (AUC = 0.876 95% CI [0.867, 0.886]) (Fig 2). The model predicting
235 hospitalization had better discrimination than the model for the need for escalation of care

236 (hospitalization: AUC = 0.829 95% CI [0.825, 0.834]; escalated hospital care: AUC = 0.805 95%
237 CI [0.795, 0.814]) (Fig 2).

238

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

Outcome	Specificity [95% CI]	Sensitivity [95% CI]	Precision [95% CI]
Hospitalization	0.73 [0.73,0.74]	0.77 [0.76,0.80]	0.74 [0.74,0.75]
Escalation of care	0.74 [0.73,0.74]	0.74 [0.72,0.76]	0.74 [0.73,0.74]
Mortality	0.78 [0.77,0.78]	0.87 [0.84,0.89]	0.79 [0.79,0.80]

241

242

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

247

248 **Model interpretation**

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

256

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

261

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

271 Fully vaccinated and boosted patients had lower predicted risks of hospitalization,
272 escalation of care, and death at 30 days. Prior SARS-CoV-2 infection predicted a lower risk of
273 mortality but a higher risk of needing hospitalization or escalation of care (Fig 3 and S1 Fig).
274 Additionally, unknown blood type and alternative insurance were among the most significant
275 predictors of a lower risk for hospitalization, while a prior diagnosis of pneumonia and no acute
276 kidney injury within two years were among the most important predictors of mortality risk (Fig
277 3A,C and S1 Fig A,C).

278 **Projected impact of risk-prioritized vaccination strategies**

279 To project the impact of targeted vaccination on adverse outcomes using the prediction
 280 models, we examined the unvaccinated subset (n=27,903) from the test cohort (n=102,859). We
 281 projected the number of adverse outcomes for three *in silico* scenarios: (1) vaccination of all
 282 Veterans within the unvaccinated subset, (2) random vaccination of 20% of the unvaccinated
 283 Veterans, and (3) vaccination of only the Veterans in the top quintile of predicted risk for adverse
 284 outcomes (Table 3). Using sensitivity tradeoff curves (data not shown), we observed a step-up of
 285 predicted risk at the top quintile. Therefore, we selected the cut-off to be the top quintile of the
 286 population. In turn, our modeling projected the optimum impact of risk-prioritized vaccination
 287 strategy. Full vaccination of the entire unvaccinated population in our test set was predicted to
 288 reduce hospitalizations by 79% (from 2,343 to 486), escalations of care by 81% (from 470 to 87),
 289 and deaths by 82% (from 167 to 30). When a random 20% of the unvaccinated population was
 290 vaccinated in the projection modeling, hospitalizations were reduced from 2343 to 2056 (12%
 291 reduction), escalations of care from 470 to 418 (11%), and deaths from 167 to 148 (11%). When
 292 vaccinating the patients in the top quintile (20%) of the highest risk for adverse outcomes,
 293 hospitalizations were reduced from 2343 to 1353 (42%), escalations of care from 470 to 309
 294 (34%), and deaths from 167 to 91 (45%).

295
 296 **Table 3: Observations and projections for occurrences for hospitalization, escalation of**
 297 **care, and mortality, for three vaccination scenarios.**

	Observed		Projections (bootstrap=100)		
<u>Outcome</u> <u>(30-day Risk)</u>	<u>Unvaccinated</u> (n=27,903)	<u>Vaccination of All</u> <u>Unvaccinated</u> [95% CI]	<u>Vaccination of Random</u> <u>20% [95% CI]</u>	<u>Vaccination in top</u> <u>Quintile (20th %ile)</u> <u>Risk [95% CI]</u>	

Hospitalization	2,343	486.17 [476.6, 495.75]	2056.06 [2052.91, 2059.21]	1353.41 [1347.3, 1359.52]
Escalation of Care	470	87.39 [84.92, 89.86]	418.37 [417.11, 419.63]	308.99 [307.40, 310.58]
Mortality	167	29.71 [28.55, 30.87]	148.23 [147.44, 149.02]	91.25 [90.37, 92.13]

298

299 Discussion

300 In a national cohort of 192,984 US Veterans who tested positive for SARS-CoV-2 during
301 the Omicron surge, we demonstrated the most robust prediction discrimination to date for 30-day
302 risk for hospitalization, escalation of care, and mortality after COVID-19 infection, using ML
303 methods. Our ML models leveraged data including detailed vaccination status and prior COVID-
304 19 infections during the Omicron surge. We identified predictors for, and projected subgroups of,
305 high-risk individuals who stand to benefit the most from advancing vaccination status. Prioritizing
306 vaccination of individuals in the highest quintile of predicted risk for hospitalization or death was
307 projected to produce greater than 3.5-fold projected reductions in hospitalization and death,
308 compared to randomly vaccinating 20% of the population.

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

317 number and timing of primary and booster vaccines, as well as prior infection, in an ML-driven
318 risk prediction model.

319 Compared to recent studies, ML models in the present study demonstrated more robust
320 discrimination by AUC in predicting 30-day risk for hospitalization (AUC 0.829), escalation of
321 care (AUC 0.805), and mortality (AUC 0.895) with COVID-19 infection. Two prior studies
322 derived from cohorts of ~4,500 patients each demonstrated lower AUCs (0.804 and 0.813) for
323 predicting hospitalization [23,24]. A previous model developed from a large VHA cohort of
324 7,635,064 (both infected and non-infected) with an observation window from May 21 to November
325 2, 2020 predicted 30-day mortality with a validation AUC of 0.836 (95% CI, 82.0%-85.3%) [9].
326 In addition, a recent study of 1,201 patients who contracted SARS-CoV-2 in Spain in 2020
327 predicted 30-day mortality with an AUC of 0.872 [25]. Commonly identified covariates in prior
328 studies, advanced age and higher medical co-morbidity indices, were associated with higher risks
329 for the adverse outcomes of interest in our models [9–11]. Our models identified a general inverse
330 association between BMI and predicted risk for adverse outcomes. This contrasts a prior meta-
331 analysis that demonstrated that higher BMI (and visceral adiposity) correlates with a higher risk
332 of hospitalization, mortality, and other adverse outcomes such as admission to ICU and need for
333 mechanical ventilation [26].

334 Consistent with prior vaccine trials [27], our study indicated that vaccination reduces
335 hospitalizations, escalation of care, and deaths. Individuals who were fully vaccinated and boosted
336 within 5 months from testing SARS-CoV-2 positive had the greatest projected protection.
337 Importantly, our model also incorporated prior COVID-19 infection as a covariate in the risk
338 modeling. Although patients with prior SARS-CoV-2 infection had a lower predicted 30-day
339 mortality, they also had higher predicted risks of 30-day hospitalization and escalated hospital

340 care. This observation is consistent with recent reports that reinfection may increase risk of any-
341 cause mortality, hospitalization, and adverse pulmonary and extra-pulmonary health outcomes
342 [28]. This enhanced risk of hospitalization and escalation of care is unclear but may be secondary
343 to attendant medical comorbidities. Use of oral anticoagulants in the two years prior to current
344 infection strongly predicted 30-day hospitalization and escalation of care. The biological basis of
345 this observation may be related to the underlying medical conditions that warranted
346 anticoagulation or to specific effects of the anticoagulants themselves. Notably, baseline
347 furosemide use was also associated a higher risk of hospitalization, escalation of care and death,
348 suggesting that underlying heart failure or volume-expanded states are important determinants of
349 infection severity in Omicron infections.

350 **Limitations**

351 The present findings in this national study of US Veterans may not be broadly applicable
352 to the general population. Consistent with the US Veteran population, our study cohort was
353 predominantly male and white with greater medical comorbidity and lower socioeconomic status
354 than the general US population. The relevance of the models remains limited for racial/ethnic
355 minority communities who have borne a disproportionate burden during the pandemic. However,
356 the methodology used here can be applied and adapted to other populations or health care systems.
357 For vaccine projections, all outcomes of interest were assumed to be the result of SARS-CoV-2
358 infection. While the VHA COVID-19 Shared Data Resource database captures all deaths, it does
359 not capture hospitalizations and care received outside the VA. This may explain why having other
360 non-VHA insurance was associated with lower rates of 30-day hospitalization given that patients
361 with non-VHA insurance may have sought care outside the VA. The VHA COVID-19 Shared Data
362 Resource database also does not establish whether SARS-CoV-2/COVID-19 is the reason for

363 hospitalization, escalation of care, or death. Determining this is challenging. Our modeling also
364 does not include laboratory or imaging data; these data have been shown to have robust predictive
365 value post index date [29–32]. Finally, the model results were most relevant to Omicron variants
366 and sub-variants and may not be relevant to other pathogenetic SARS-CoV-2 variants.

367 **Conclusions**

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

377 **Acknowledgements**

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

380

381

382

383 **References**

- 384 1. WHO Coronavirus (COVID-19) Dashboard. [cited 2 Jun 2022]. Available:
385 <https://covid19.who.int>
- 386 2. CDC COVID Data Tracker. [cited 18 Mar 2022]. Available: [https://covid.cdc.gov/covid-](https://covid.cdc.gov/covid-data-tracker/#variant-proportions)
387 [data-tracker/#variant-proportions](https://covid.cdc.gov/covid-data-tracker/#variant-proportions)
- 388 3. Iuliano AD, Brunkard JM, Boehmer TK, Peterson E, Adjei S, Binder AM, et al. Trends in
389 Disease Severity and Health Care Utilization During the Early Omicron Variant Period
390 Compared with Previous SARS-CoV-2 High Transmission Periods - United States,
391 December 2020-January 2022. *MMWR Morb Mortal Wkly Rep.* 2022;71: 146–152.
392 doi:10.15585/mmwr.mm7104e4
- 393 4. Classification of Omicron (B.1.1.529): SARS-CoV-2 Variant of Concern. [cited 8 Jun
394 2022]. Available: [https://www.who.int/news/item/26-11-2021-classification-of-omicron-](https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern)
395 [\(b.1.1.529\)-sars-cov-2-variant-of-concern](https://www.who.int/news/item/26-11-2021-classification-of-omicron-(b.1.1.529)-sars-cov-2-variant-of-concern)
- 396 5. VanBlargan LA, Errico JM, Halfmann PJ, Zost SJ, Crowe JE, Purcell LA, et al. An
397 infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic
398 monoclonal antibodies. *Nat Med.* 2022;28: 490–495. doi:10.1038/s41591-021-01678-y
- 399 6. Abdullah F, Myers J, Basu D, Tintinger G, Ueckermann V, Mathebula M, et al. Decreased
400 severity of disease during the first global omicron variant covid-19 outbreak in a large
401 hospital in tshwane, south africa. *Int J Infect Dis IJID Off Publ Int Soc Infect Dis.*
402 2022;116: 38–42. doi:10.1016/j.ijid.2021.12.357
- 403 7. Chen J, Wang R, Gilby NB, Wei G-W. Omicron Variant (B.1.1.529): Infectivity, Vaccine
404 Breakthrough, and Antibody Resistance. *J Chem Inf Model.* 2022;62: 412–422.
405 doi:10.1021/acs.jcim.1c01451
- 406 8. Nyberg T, Ferguson NM, Nash SG, Webster HH, Flaxman S, Andrews N, et al.
407 Comparative analysis of the risks of hospitalisation and death associated with SARS-CoV-2
408 omicron (B.1.1.529) and delta (B.1.617.2) variants in England: a cohort study. *The Lancet.*
409 2022;399: 1303–1312. doi:10.1016/S0140-6736(22)00462-7
- 410 9. Ioannou GN, Green P, Fan VS, Dominitz JA, O’Hare AM, Backus LI, et al. Development
411 of COVIDVax Model to Estimate the Risk of SARS-CoV-2–Related Death Among 7.6
412 Million US Veterans for Use in Vaccination Prioritization. *JAMA Netw Open.* 2021;4:
413 e214347. doi:10.1001/jamanetworkopen.2021.4347
- 414 10. Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P, et al. Prediction for Progression Risk in
415 Patients With COVID-19 Pneumonia: The CALL Score. *Clin Infect Dis Off Publ Infect Dis*
416 *Soc Am.* 2020;71: 1393–1399. doi:10.1093/cid/ciaa414
- 417 11. Jung C, Excoffier J-B, Raphaël-Rousseau M, Salaün-Penquer N, Ortala M, Chouaid C.
418 Evolution of hospitalized patient characteristics through the first three COVID-19 waves in

- 419 Paris area using machine learning analysis. *PloS One*. 2022;17: e0263266.
420 doi:10.1371/journal.pone.0263266
- 421 12. Liang W, Liang H, Ou L, Chen B, Chen A, Li C, et al. Development and Validation of a
422 Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients
423 With COVID-19. *JAMA Intern Med*. 2020;180: 1081–1089.
424 doi:10.1001/jamainternmed.2020.2033
- 425 13. UpSetR: an R package for the visualization of intersecting sets and their properties |
426 Bioinformatics | Oxford Academic. [cited 14 Jul 2022]. Available:
427 <https://academic.oup.com/bioinformatics/article/33/18/2938/3884387>
- 428 14. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic
429 comorbidity in longitudinal studies: development and validation. *J Chronic Dis*. 1987;40:
430 373–383. doi:10.1016/0021-9681(87)90171-8
- 431 15. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity Measures for Use with
432 Administrative Data. *Med Care*. 1998;36: 8–27.
- 433 16. LeDell E, Poirier S. H2O AutoML: Scalable Automatic Machine Learning. 7th ICML
434 Workshop Autom Mach Learn AutoML. 2020. Available: [https://www.automl.org/wp-](https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf)
435 [content/uploads/2020/07/AutoML_2020_paper_61.pdf](https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf)
- 436 17. mateuszbuda. Machine Learning Statistical Utils. 2022. Available:
437 <https://github.com/mateuszbuda/ml-stat-util>
- 438 18. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn:
439 Machine Learning in Python. *J Mach Learn Res*. 2011;12: 2825–2830.
- 440 19. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. *arXiv*; 2017
441 Nov. Report No.: arXiv:1705.07874. doi:10.48550/arXiv.1705.07874
- 442 20. Andrews N, Stowe J, Kirsebom F, Toffa S, Rickeard T, Gallagher E, et al. Covid-19
443 Vaccine Effectiveness against the Omicron (B.1.1.529) Variant. *N Engl J Med*. 2022;386:
444 1532–1546. doi:10.1056/NEJMoa2119451
- 445 21. Kelly JD, Leonard S, Hoggatt KJ, Boscardin WJ, Lum EN, Moss-Vazquez TA, et al.
446 Incidence of Severe COVID-19 Illness Following Vaccination and Booster With
447 BNT162b2, mRNA-1273, and Ad26.COV2.S Vaccines. *JAMA*. 2022;328: 1427–1437.
448 doi:10.1001/jama.2022.17985
- 449 22. Ong SWX, Tham SM, Koh LP, Dugan C, Khoo BY, Ren D, et al. External validation of the
450 PRIORITY model in predicting COVID-19 critical illness in vaccinated and unvaccinated
451 patients. *Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis*. 2022;28:
452 884.e1-884.e3. doi:10.1016/j.cmi.2022.01.031

- 453 23. Jehi L, Ji X, Milinovich A, Erzurum S, Merlino A, Gordon S, et al. Development and
454 validation of a model for individualized prediction of hospitalization risk in 4,536 patients
455 with COVID-19. *PLOS ONE*. 2020;15: e0237419. doi:10.1371/journal.pone.0237419
- 456 24. Willette AA, Willette SA, Wang Q, Pappas C, Klinedinst BS, Le S, et al. Using machine
457 learning to predict COVID-19 infection and severity risk among 4510 aged adults: a UK
458 Biobank cohort study. *Sci Rep*. 2022;12: 7736. doi:10.1038/s41598-022-07307-z
- 459 25. Reina Reina A, Barrera JM, Valdivieso B, Gas M-E, Maté A, Trujillo JC. Machine learning
460 model from a Spanish cohort for prediction of SARS-COV-2 mortality risk and critical
461 patients. *Sci Rep*. 2022;12: 5723. doi:10.1038/s41598-022-09613-y
- 462 26. Demeulemeester F, de Punder K, van Heijningen M, van Doesburg F. Obesity as a Risk
463 Factor for Severe COVID-19 and Complications: A Review. *Cells*. 2021;10: 933.
464 doi:10.3390/cells10040933
- 465 27. Polack FP, Thomas SJ, Kitchin N, Absalon J, Gurtman A, Lockhart S, et al. Safety and
466 Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *N Engl J Med*. 2020;383: 2603–
467 2615. doi:10.1056/NEJMoa2034577
- 468 28. Bowe B, Xie Y, Al-Aly Z. Acute and postacute sequelae associated with SARS-CoV-2
469 reinfection. *Nat Med*. 2022;28: 2398–2405. doi:10.1038/s41591-022-02051-3
- 470 29. Alle S, Kanakan A, Siddiqui S, Garg A, Karthikeyan A, Mehta P, et al. COVID-19 Risk
471 Stratification and Mortality Prediction in Hospitalized Indian Patients: Harnessing clinical
472 data for public health benefits. *PLoS ONE*. 2022;17: e0264785.
473 doi:10.1371/journal.pone.0264785
- 474 30. Butler L, Karabayir I, Samie Tootooni M, Afshar M, Goldberg A, Akbilgic O. Image and
475 structured data analysis for prognostication of health outcomes in patients presenting to the
476 ED during the COVID-19 pandemic. *Int J Med Inf*. 2021;158: 104662.
477 doi:10.1016/j.ijmedinf.2021.104662
- 478 31. Ortiz A, Trivedi A, Desbiens J, Blazes M, Robinson C, Gupta S, et al. Effective deep
479 learning approaches for predicting COVID-19 outcomes from chest computed tomography
480 volumes. *Sci Rep*. 2022;12: 1716. doi:10.1038/s41598-022-05532-0
- 481 32. Zhang K, Liu X, Shen J, Li Z, Sang Y, Wu X, et al. Clinically Applicable AI System for
482 Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia
483 Using Computed Tomography. *Cell*. 2020;181: 1423-1433.e11.
484 doi:10.1016/j.cell.2020.04.045

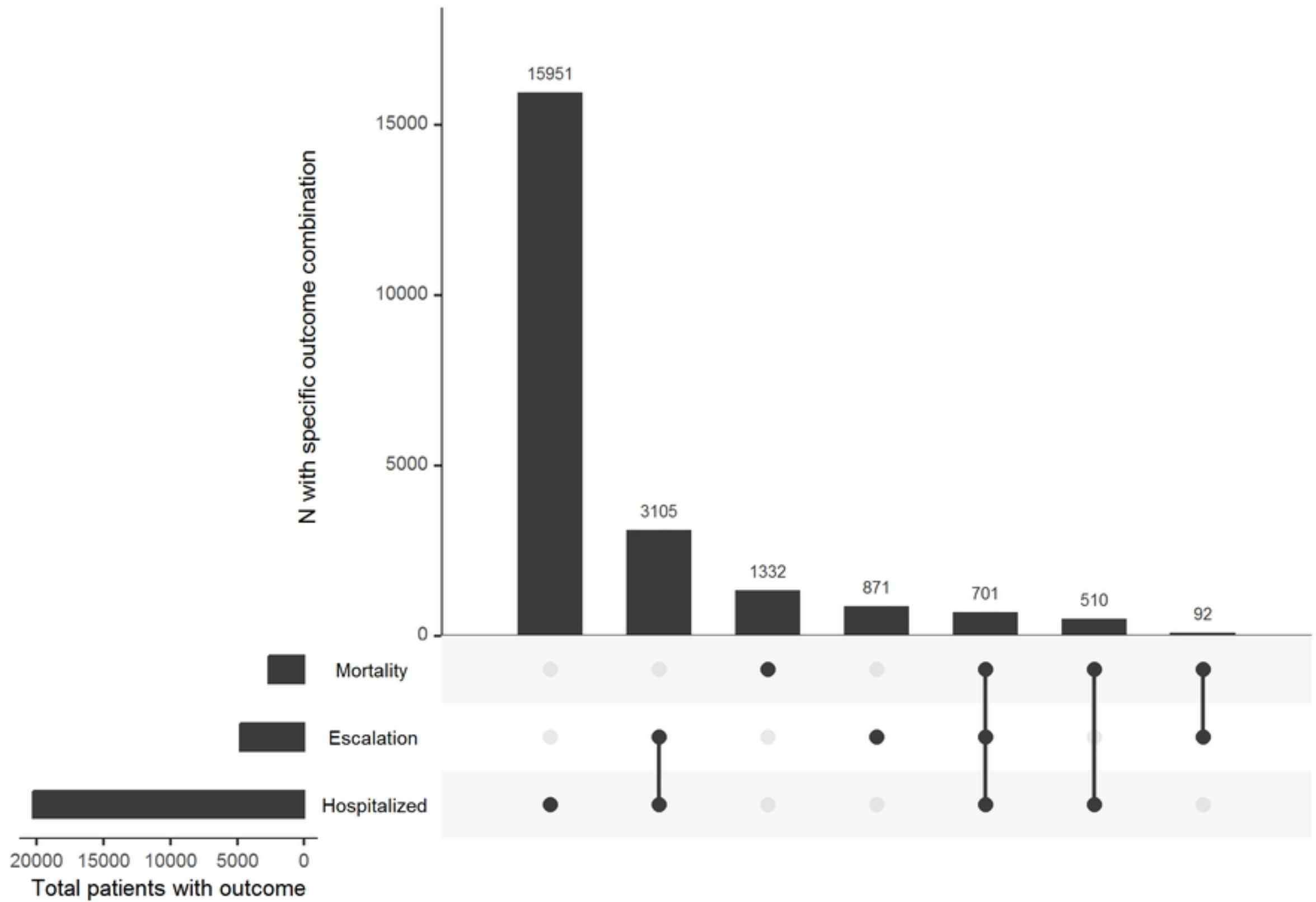
485

486 **Supporting Information**

487 **S1 Table. Covariates used in predictive modeling.** A table of all potential covariates that were
488 investigated with a brief definition.

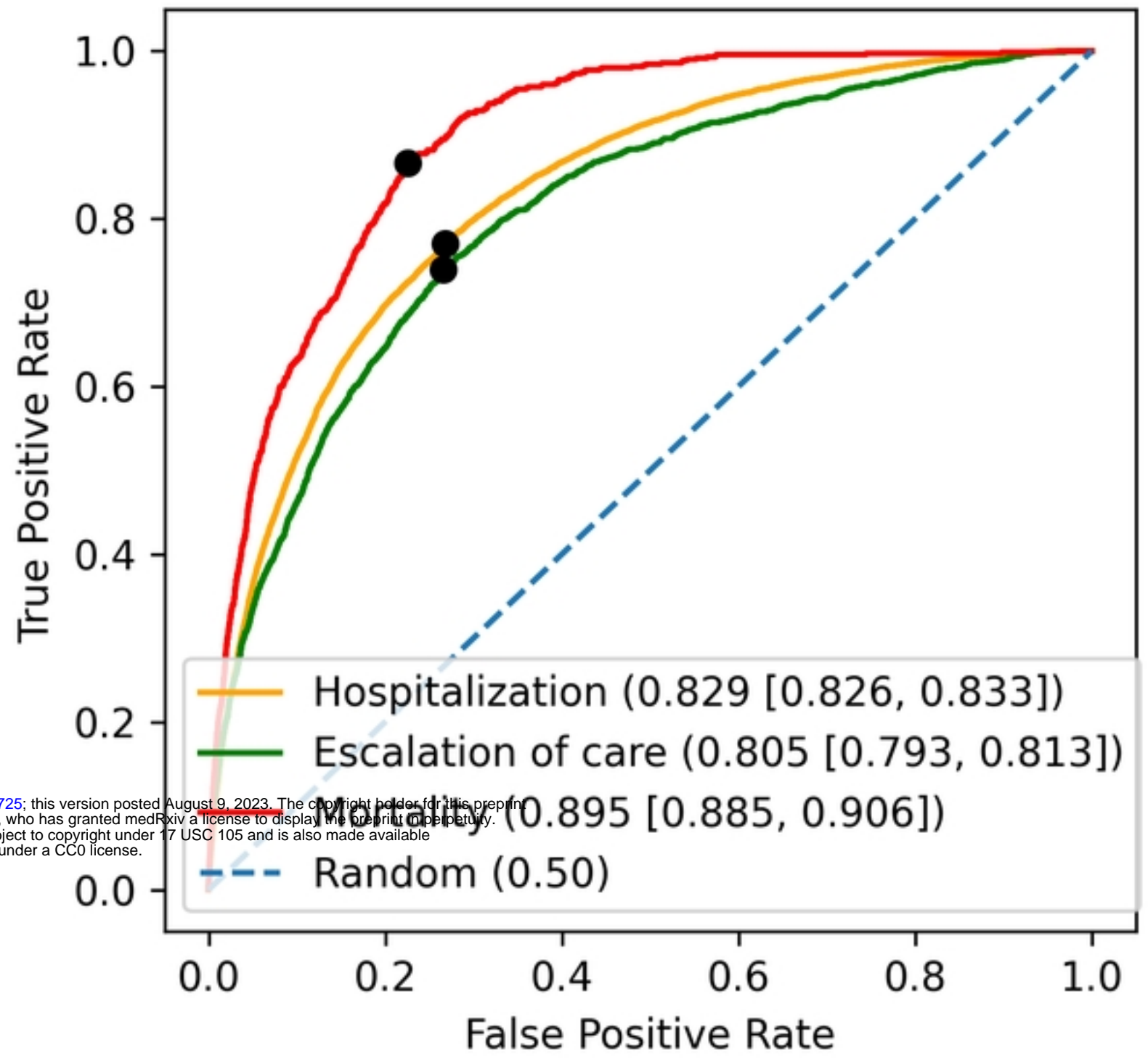
489 **S1 Fig. SHAP summary plots for 30-day outcomes of interest.** (A) hospitalization, (B)
490 escalation of care, and (C) mortality. Covariates are listed in order of highest to lowest impact
491 (based on absolute mean SHAP value) along the y-axis. Each blue or red point represents a
492 patient's specified covariate value; that value is color coded in a heat map fashion per the legend.
493 The x-axis is the SHAP value for the specific covariate, with SHAP values greater than 0 indicating
494 higher predicted risk contribution and values less than 0 indicating lower predicted risk
495 contribution for the given outcome.

496



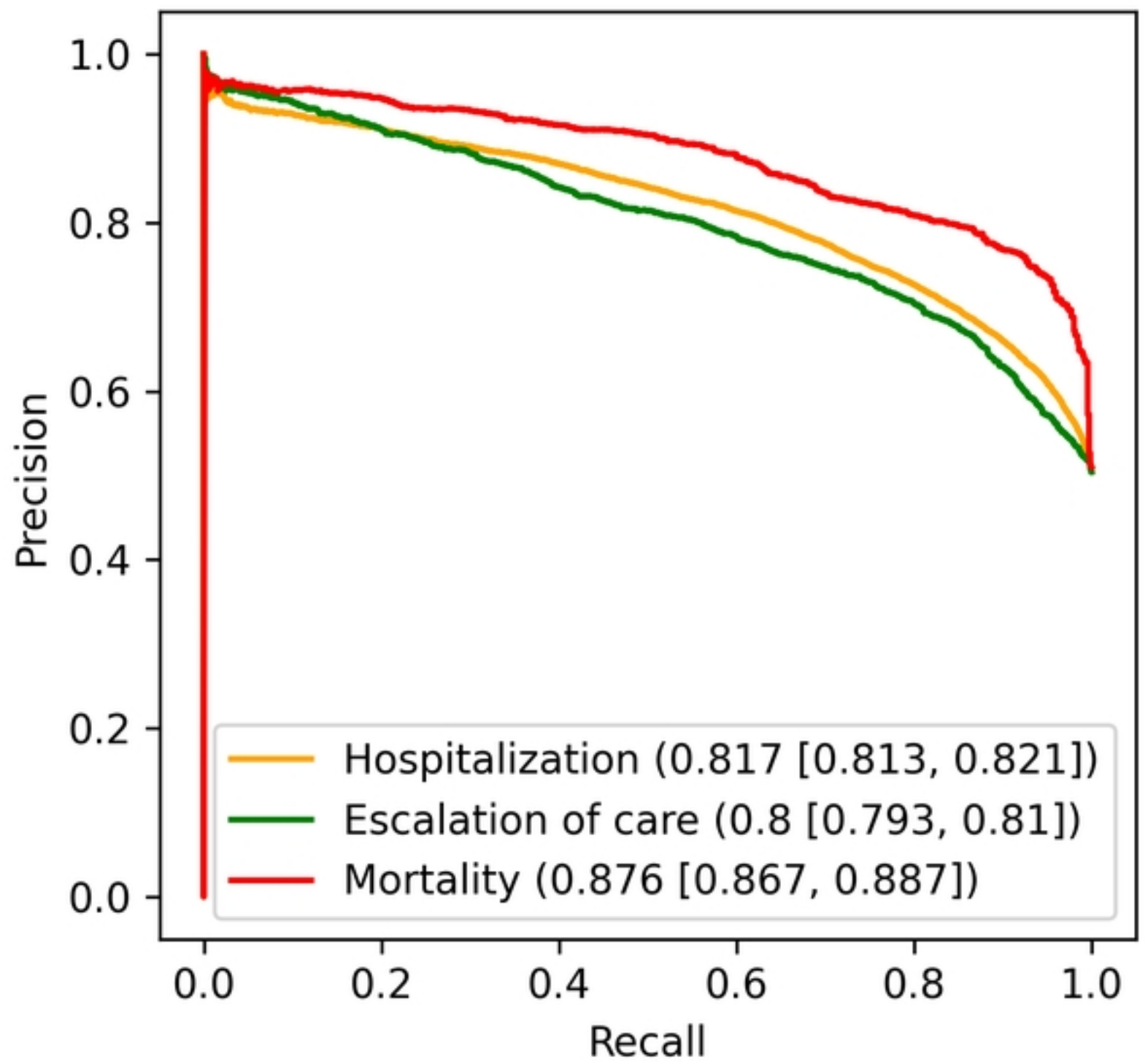
Figure

(A)

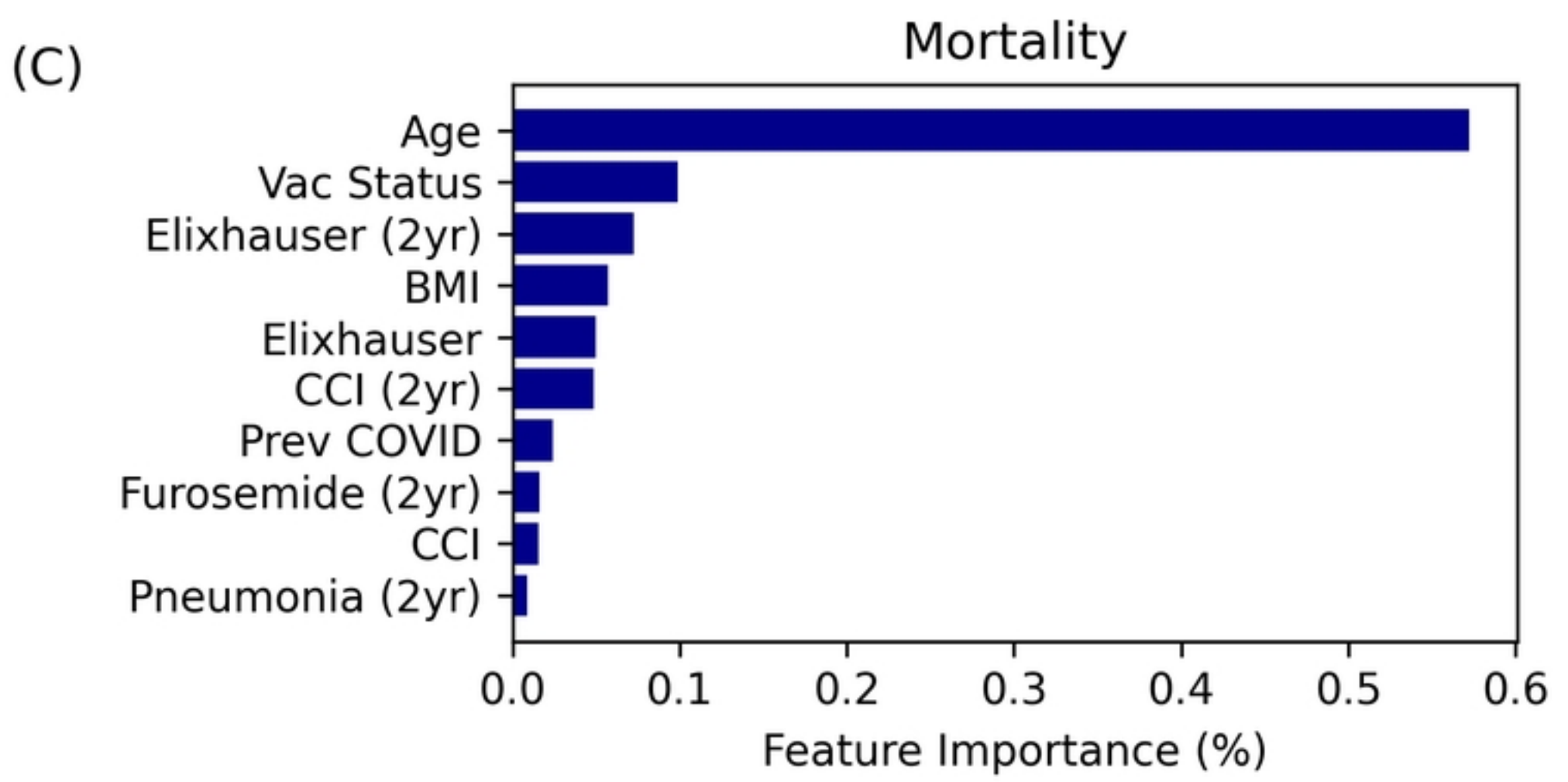
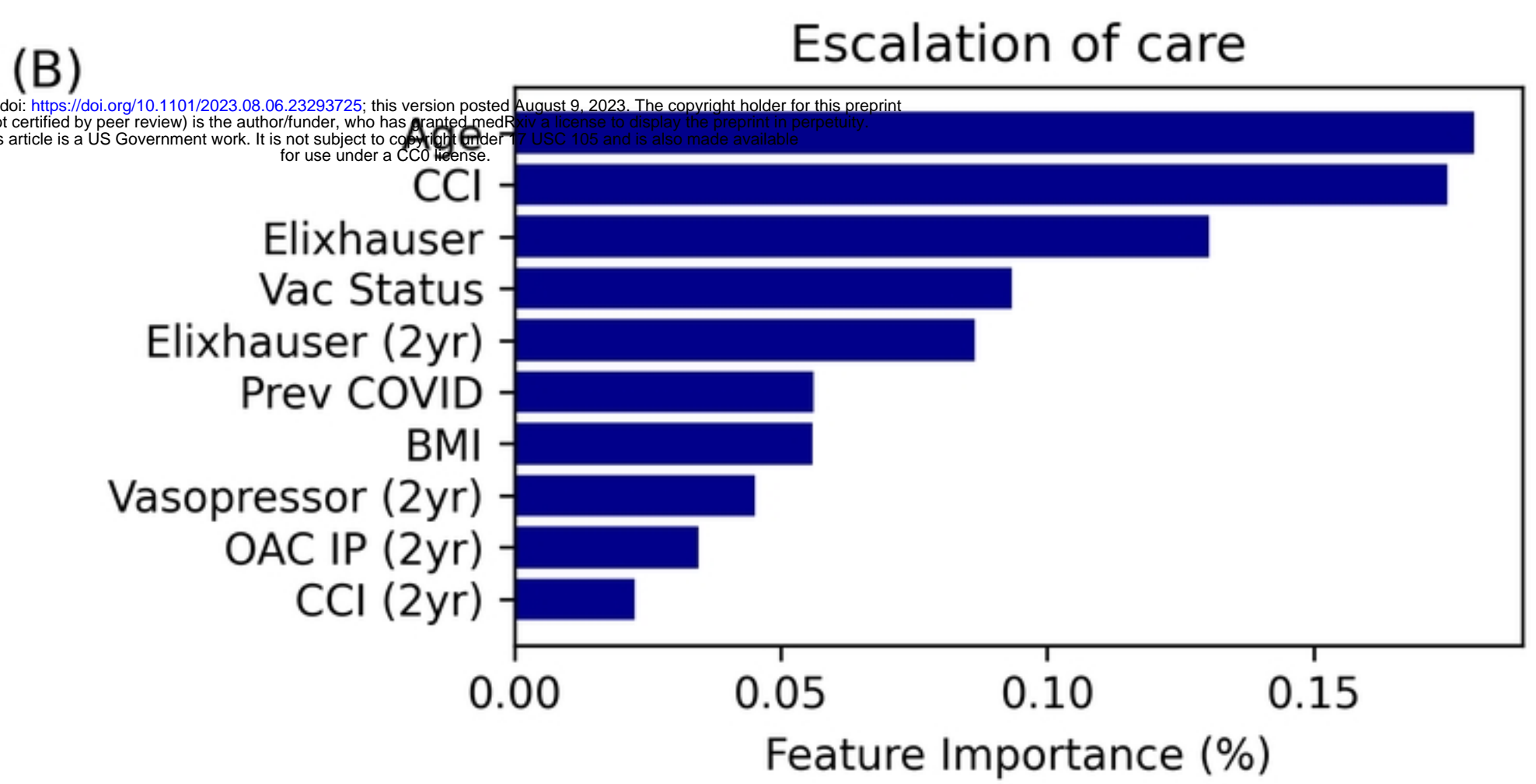
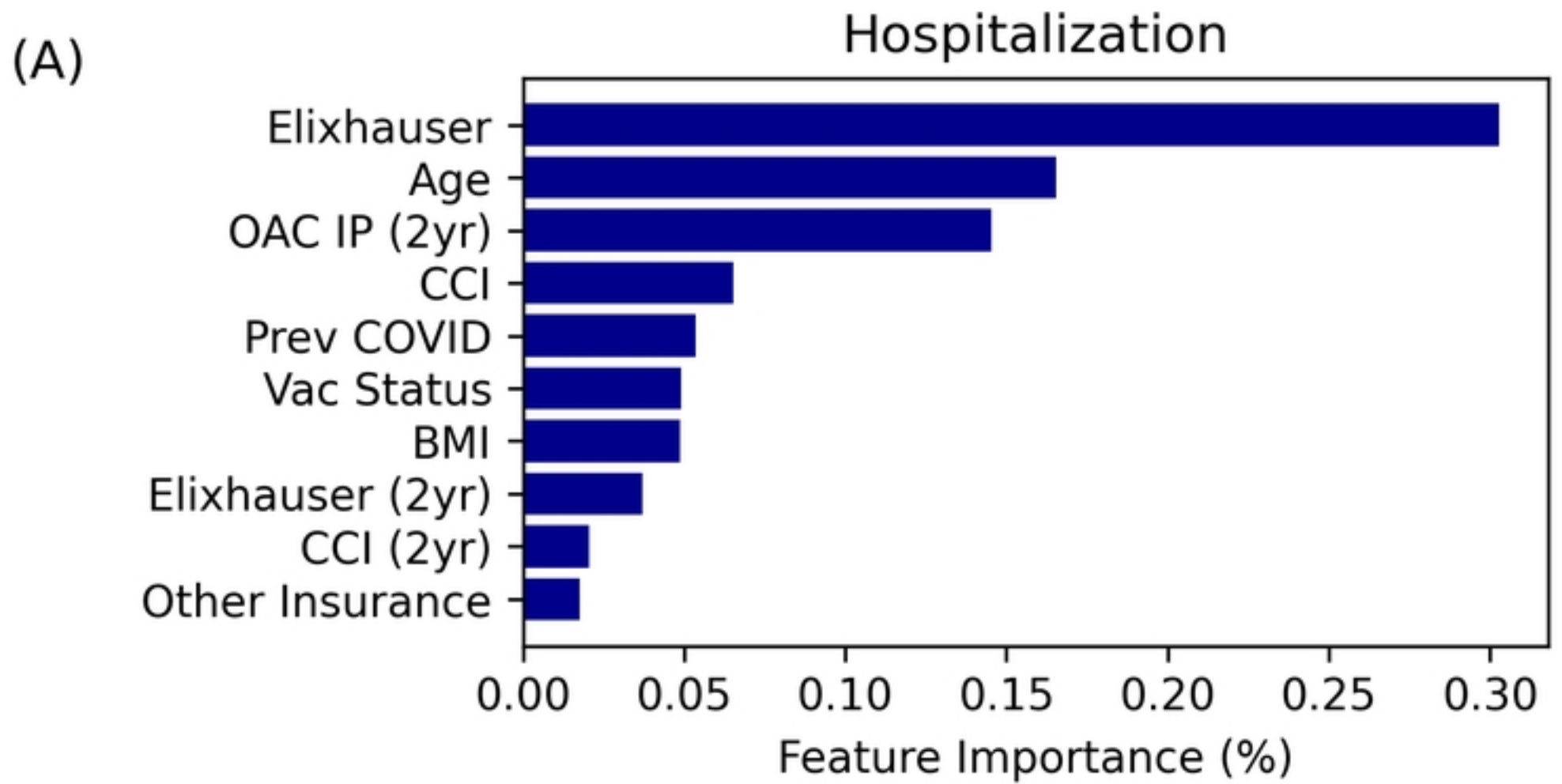


medRxiv preprint doi: <https://doi.org/10.1101/2023.08.06.23293725>; this version posted August 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

(B)

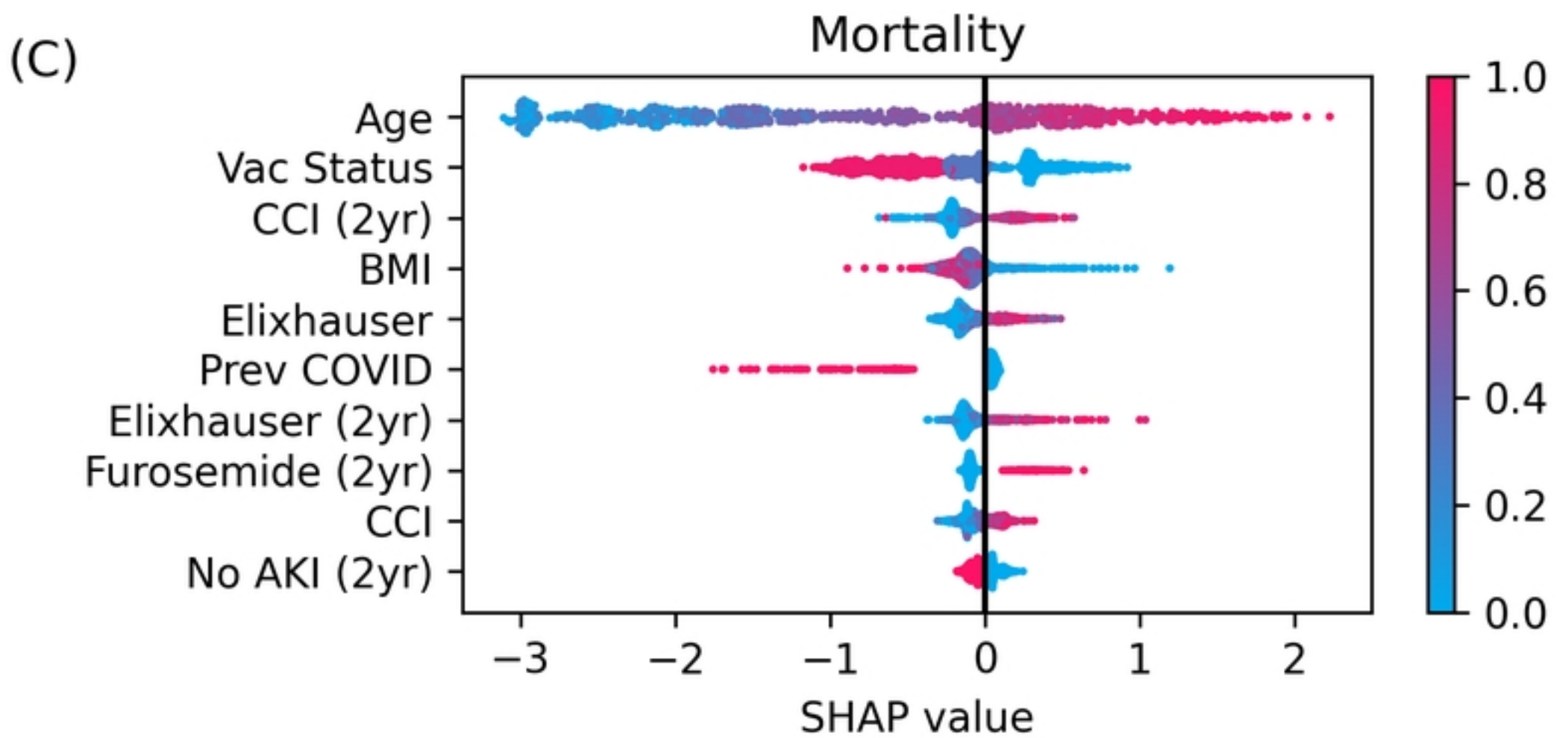
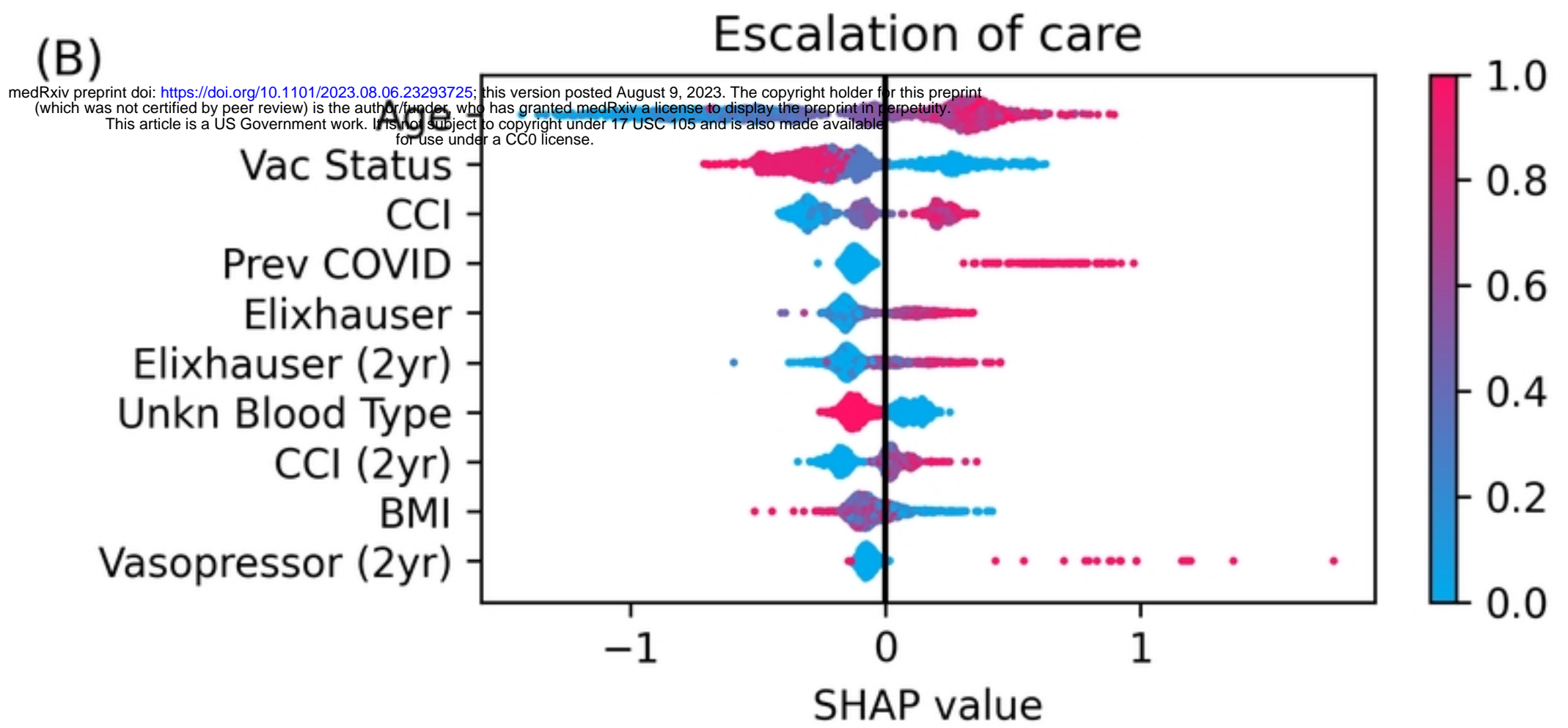
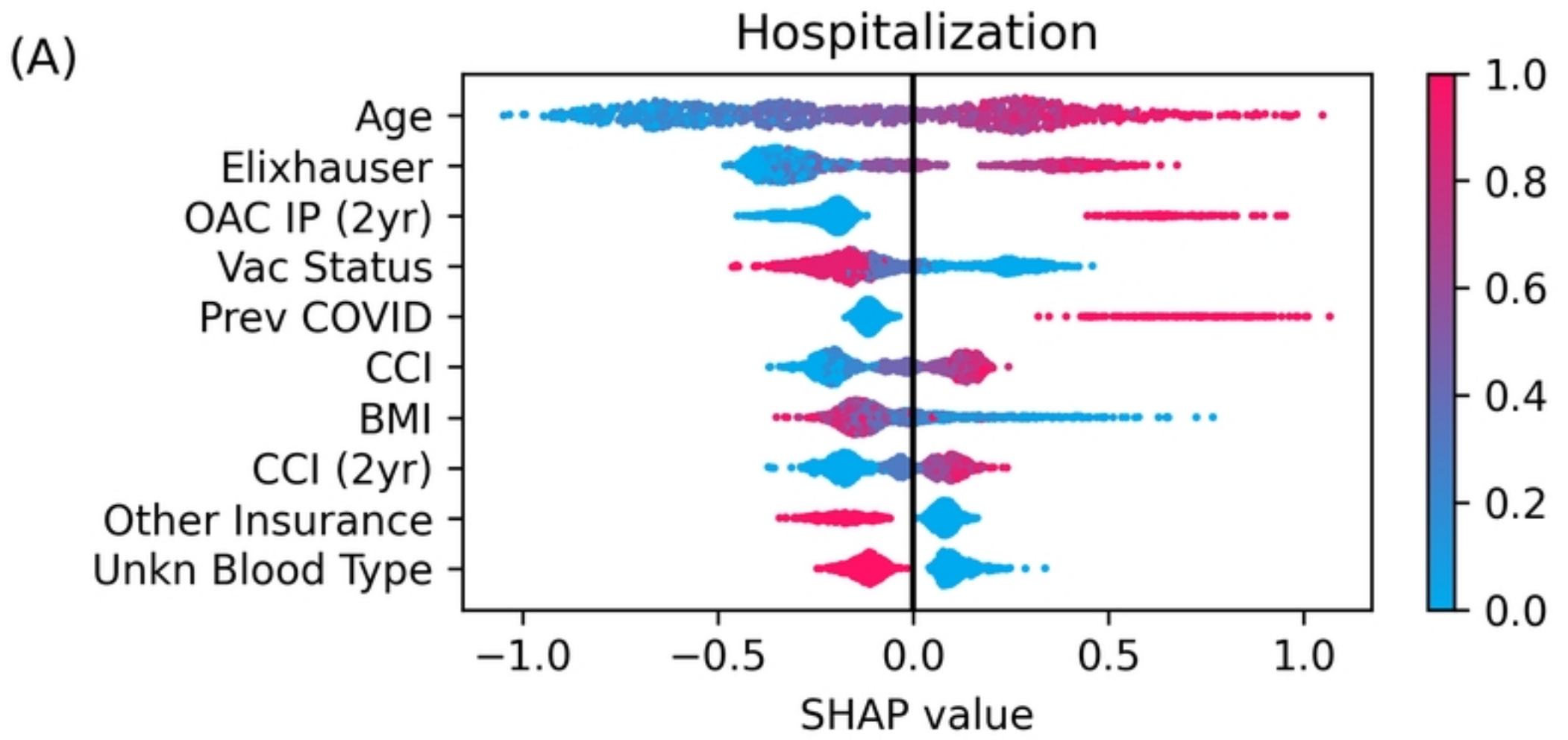


Figure



medRxiv preprint doi: <https://doi.org/10.1101/2023.08.06.23293725>; this version posted August 9, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. This article is a US Government work. It is not subject to copyright under 17 USC 105 and is also made available for use under a CC0 license.

Figure



Figure