

1 **Title: Comparing Machine Learning Algorithms for Predicting ICU**

2 **Admission and Mortality in COVID-19**

3 **Authors:** Sonu Subudhi¹, Ashish Verma^{2*}, Ankit B. Patel^{2*}, C. Corey Hardin³, Melin J.
4 Khandekar⁴, Hang Lee⁵, Triantafyllos Stylianopoulos⁶, Lance L. Munn⁷, Sayon Dutta^{8#} and
5 Rakesh K. Jain^{7#}

6 **Affiliations:**

7 ¹Department of Medicine/Gastroenterology Division, Massachusetts General Hospital and
8 Harvard Medical School, Boston, Massachusetts

9 ²Department of Medicine/Renal Division, Brigham and Women's Hospital and Harvard Medical
10 School, Boston, Massachusetts

11 ³Department of Pulmonary and Critical Care Medicine, Massachusetts General Hospital and
12 Harvard Medical School, Boston, Massachusetts

13 ⁴Department of Radiation Oncology, Massachusetts General Hospital and Harvard Medical
14 School, Boston, Massachusetts

15 ⁵Biostatistics Center, Massachusetts General Hospital and Harvard Medical School, Boston,
16 Massachusetts

17 ⁶Cancer Biophysics Laboratory, Department of Mechanical and Manufacturing Engineering,
18 University of Cyprus, Nicosia, Cyprus

19 ⁷Edwin L. Steele Laboratories, Department of Radiation Oncology, Massachusetts General
20 Hospital and Harvard Medical School, Boston, Massachusetts

21 ⁸Department of Emergency Medicine, Massachusetts General Hospital and Harvard Medical
22 School, Boston, Massachusetts

23 *Equal contribution

24 [#]To whom correspondence should be addressed: Sayon Dutta, MD (sdutta1@partners.org) and
25 Rakesh K. Jain, PhD (jain@steele.mgh.harvard.edu)

26 **Short title:** Comparing machine learning algorithms in COVID-19

27

28 **Abstract (150 words):** As predicting the trajectory of COVID-19 disease is challenging,
29 machine learning models could assist physicians determine high-risk individuals. This study
30 compares the performance of 18 machine learning algorithms for predicting ICU admission and
31 mortality among COVID-19 patients. Using COVID-19 patient data from the Mass General
32 Brigham (MGB) healthcare database, we developed and internally validated models using
33 patients presenting to Emergency Department (ED) between March-April 2020 (n = 1144) and
34 externally validated them using those individuals who encountered ED between May-August
35 2020 (n = 334). We show that ensemble-based models perform better than other model types at
36 predicting both 5-day ICU admission and 28-day mortality from COVID-19. CRP, LDH, and
37 procalcitonin levels were important for ICU admission models whereas eGFR <60
38 ml/min/1.73m², ventilator use, and potassium levels were the most important variables for
39 predicting mortality. Implementing such models would help in clinical decision-making for
40 future COVID-19 and other infectious disease outbreaks.

41 **[Main Text: 3422 words]**

42 **Introduction**

43 The COVID-19 pandemic has led to significant morbidity and mortality throughout the
44 world ¹. The rapid spread of SARS-CoV-2 has provided limited time to identify factors involved
45 in SARS-CoV-2 transmission, predictors of COVID-19 severity, and effective treatments. At the
46 height of the pandemic, areas with high number of SARS-CoV-2 infections were resource-
47 limited and forced to ration life-saving therapies such as ventilators and dialysis machines ^{2,3}. In
48 this setting, being able to identify patients requiring intensive care or at high risk of mortality
49 upon presentation to the hospital may help providers expedite patients to the most appropriate
50 care setting.

51 Model predictions are gaining increasing interest in clinical medicine. Machine learning
52 applications have been used to help predict acute kidney injury ⁴ and septic shock ⁵, amongst
53 other outcomes in hospitalized patients. These tools have also been applied to outpatients to
54 predict outcomes such as heart failure progression ⁶. Machine learning tools can be applied to
55 predict outcomes such as Intensive Care Unit (ICU) admission and mortality ⁷. Thus far there
56 have been few studies that examined specific machine learning algorithms in predicting
57 outcomes such as mortality in COVID-19 patients ⁸⁻¹⁰. Given the potential utility of machine
58 learning-based decision rules and the urgency of the pandemic, a concerted effort is being made
59 to identify which machine learning applications are optimal for given sets of data and diseases ¹¹.
60 To address this knowledge gap, we conducted a multi-hospital cohort (Mass General Brigham
61 (MGB) healthcare database) study to extensively evaluate the performance of 18 different
62 machine learning algorithms in predicting ICU admission and mortality. Our goal was to identify
63 the best prognostication algorithm using demographic data, comorbidities, and laboratory
64 findings of COVID-19 patients who visited emergency departments (ED) at MGB between
65 March and April 2020. We validated our models on a temporally distinct patient cohort that

66 tested positive for COVID-19 and had ED encounter between May and August 2020. We also
67 identified critical variables utilized by the model to predict ICU admission and mortality.

68 **Results**

69 *Patient characteristics.*

70 We obtained data from 10,826 patients in the multihospital database (Massachusetts
71 General Brigham Healthcare database) who had COVID-19 infection during the period of March
72 and April 2020. A total of 3,713 out of the 10,826 patients visited EDs. We evaluated patients
73 based on demographics, medication use, history of past illness, clinical features, and laboratory
74 values described in Table S1. After excluding patients with missing data, 1,144 patients
75 remained, 99% of which were in-patients (n = 1133). For external validation, we pulled data of
76 temporally distinct individuals from the Mass General Brigham (MGB) healthcare database who
77 were positive for SARS-CoV-2 between May and August 2020. During this period, 1,754 out of
78 8,013 SARS-CoV-2 positive individuals visited the ER. After excluding patients with missing
79 variables from Table S1, a total of 334 patients were left (98% of which were in-patients).

80 The baseline characteristics of 1,144 patients in the training dataset are listed in Table 1.
81 The overall study population included 45% women, and the majority were above the age of 60.
82 The number of patients who were admitted to ICU within 5 days and who died within 28 days of
83 ED visit were 342 (30%) and 217 (19%), respectively. The external validation dataset included
84 patients with similar distribution in age ≥ 50 years ($X^2_{(4, N = 1193)} = 8.9, p = 0.063$), gender ($X^2_{(1, N =$
85 $1478)} = 0.017, p = 0.89$), race ($X^2_{(1, N = 1478)} = 0.07, p = 0.79$) and BMI ($X^2_{(2, N = 1478)} = 4.31, p =$
86 0.12) (Table S6). Of the 334 patients who visited the ED, 74 (22%) were admitted to the ICU
87 and 45 (13%) died with COVID-19.

88 ***Comparing performance of prediction models – cross validation.***

89 We evaluated 18 machine learning algorithms belonging to 9 broad categories, namely
90 ensemble, Gaussian process, linear, naïve bayes, nearest neighbor, support vector machine, tree-
91 based, discriminant analysis and neural network models.

92 On comparing the ICU admission prediction models using cross validation, we observed
93 that all ensemble-based models had mean precision-recall area under curve (PR AUC) scores
94 more than 0.77 (Table 2; Fig. S2A-B). Specifically, the PR AUC score for *AdaBoostClassifier*
95 was 0.80 (95% CI, 0.73 – 0.87), for *BaggingClassifier* was 0.80 (95% CI, 0.73 – 0.87), for
96 *GradientBoostingClassifier* was 0.77 (95% CI, 0.68 – 0.86), for *RandomForestClassifier* was
97 0.80 (95% CI, 0.70 – 0.90), for *XGBClassifier* was 0.78 (95% CI, 0.70 – 0.86), and for
98 *ExtraTreesClassifier* was [0.79 (95% CI, 0.72 – 0.86)]. In addition, *LogisticRegression* [0.79
99 (95% CI, 0.71 – 0.87)], and *LinearDiscriminantAnalysis* [0.76 (95% CI, 0.68 – 0.84)] also had
100 high PR AUC scores. In contrast, *GaussianProcessClassifier* [0.6 (95% CI, 0.54 – 0.66)],
101 *SGDClassifier* [0.63 (95% CI, 0.60 – 0.66)] and *LinearSVC* [0.65 (95% CI, 0.57 – 0.73)] had
102 low PR AUC scores. Upon performing multiple comparison analysis between all models (based
103 on PR AUC and ROC AUC scores), the ensemble-based models and *LogisticRegression* models
104 have similar pattern of performance (Fig. S1A-B). On grouping the models based on their broad
105 categories, we found that ensemble models have significantly higher PR AUC scores than all
106 other model types except for logistic regression (based on Fisher's Least Significant Difference
107 (LSD) t-test; Fig. 2A; details of statistical analysis in Table S7). For ROC AUC scores, ensemble
108 models performed better than all models except logistic regression (Fig. 2A; Table S7).

109 On comparing the mortality prediction models using cross validation, all ensemble-based
110 models had mean PR AUC scores higher than 0.8 (Table 3; Fig. S2C-D). The PR AUC score for

111 *AdaBoostClassifier* was 0.81 (95% CI, 0.76 – 0.86), for *BaggingClassifier* was 0.81 (95% CI,
112 0.74 – 0.88), for *GradientBoostingClassifier* was 0.81 (95% CI, 0.73 – 0.89), for
113 *RandomForestClassifier* was 0.8 (95% CI, 0.75 – 0.85), for *XGBClassifier* was 0.82 (95% CI,
114 0.75 – 0.89), and *ExtraTreesClassifier* [0.82 (95% CI, 0.74 – 0.90)]. In addition,
115 *LinearDiscriminantAnalysis* [0.85 (95% CI, 0.79 – 0.91)] also had a high PR AUC score.
116 However, for mortality prediction, *LogisticRegression* [0.73 (95% CI, 0.62 – 0.84)] had low PR
117 AUC score when compared to ensemble methods. The lowest PR AUC scores were for
118 *GaussianProcessClassifier* [0.55 (95% CI, 0.42 – 0.68)], *SGDClassifier* [0.54 (95% CI, 0.49 –
119 0.59)], *Perceptron* [0.6 (95% CI, 0.53 – 0.67)], and *KNeighborsClassifier* [0.6 (95% CI, 0.52 –
120 0.68)]. Upon performing multiple comparison analysis between all models (based on PR AUC
121 and ROC AUC scores), the ensemble-based models and *LinearDiscriminantAnalysis* models had
122 similar patterns of performance (Fig. S1C-D). When we grouped the models based on their broad
123 categories and compared their PR AUC and ROC AUC scores, we found that ensemble-based
124 models perform better than all other model types except Naïve bayes and discriminant analysis
125 based methods (based on Fisher's Least Significant Difference (LSD) t-test; Fig. 2B; details of
126 statistical analysis in Table S7).

127 ***Comparing performance of prediction models – internal and external validation.***

128 We then tested the internal validation dataset on ICU admission models and found that
129 ensemble methods (PR AUC \geq 0.8) and *LogisticRegression* (PR AUC = 0.83) had the best scores
130 (Table 2). However, for the external validation dataset, *BaggingClassifier*,
131 *RandomForestClassifier* and *XGBClassifier* had better PR AUC scores (\geq 0.6) than other
132 ensemble models. *LogisticRegression* also performed comparably (PR AUC = 0.62) to well-
133 performing ensemble methods with the external validation dataset.

134 On evaluating the performance of mortality models using internal validation dataset,
135 ensemble methods, naïve bayes, and discriminant analysis-based models outperformed other
136 models (PR AUC \geq 0.7) (Table 3). In the external validation dataset, although the PR AUC
137 scores were lower, *AdaBoostClassifier*, *BaggingClassifier*, and *RandomForestClassifier* had
138 better PR AUC scores (\geq 0.37) than other models. Unlike ICU admission prediction,
139 *LogisticRegression* had a low score with internal and external validation datasets (PR AUC =
140 0.65 and 0.23, respectively).

141 Overall, we found that ensemble models performed well in predicting both ICU
142 admission and mortality for COVID-19 patients.

143 ***Critical variables for predicting ICU admission and mortality.***

144 To investigate how individual variables in the machine learning models impact outcome
145 prediction, we performed SHAP analysis for the best models – namely random forest for the ICU
146 admission model and XGB classifier for the mortality prediction model. For the ICU admission
147 prediction models, C-reactive protein, procalcitonin, lactate dehydrogenase, and first respiratory
148 rate were directly proportional to risk of ICU admission (Fig. 2C-D), while lower values of the
149 first oxygen saturation reading and lymphocytes were associated with increased probability of
150 ICU admission. For mortality prediction models, use of ventilator, estimated glomerular filtration
151 rate less than 60 ml/min/1.72 m², age greater than 80 years, hyperkalemia and high procalcitonin
152 were associated with higher mortality while lower lymphocyte counts were associated with
153 increased probability of death (Fig. 2E-F).

154 **Discussion**

155 In this study, we evaluated the ability of various machine learning algorithms to predict
156 clinical outcomes such as ICU admission or mortality using data available from initial ER
157 encounter of COVID-19 patients. Based on our analysis of 18 algorithms, we found that
158 ensemble-based methods have moderately better performance than other machine learning
159 algorithms. Optimizing the hyperparameters (Tables S4 and S5) enabled us to achieve the best-
160 performing ensemble models. We also identified variables that had the largest impact on the
161 performance of the models. We demonstrated that for predicting ICU admission, C-reactive
162 protein, LDH, procalcitonin, lymphocytes, neutrophils, oxygen saturation and respiratory rate are
163 among the top predictors, but for mortality prediction, eGFR < 60 ml/min/1.73m², serum
164 potassium levels, use of ventilator, age, ALT and white blood cells are the leading predictors.

165 Our model detected that CRP, LDH, procalcitonin, eGFR < 60 ml/min/m², serum
166 potassium levels, advanced age and ventilator use are indicative of a worse outcome, which
167 aligns with previous studies of ICU admission and mortality (Table S2). Multiple retrospective
168 studies showed that increased procalcitonin values were associated with high risk for severe
169 COVID-19 infection¹². The explanation behind this association is not clear. Increased
170 procalcitonin level in COVID -19 patients can suggest bacterial coinfection, a marker of severity
171 of ARDS and immune dysregulation¹³⁻¹⁵ but may also be a marker of the hyperinflammation
172 associated with COVID-19 severity. We also found reduced kidney function as the major risk
173 factor for ICU mortality. This result has been revealed by two previous studies in the literature,
174 indicating that patients on dialysis and with chronic kidney disease have a high risk of mortality
175 from COVID-19^{16,17}. Our study also highlighted serum potassium level as an important predictor
176 for mortality. This finding has not been reported in the literature to our knowledge, although one
177 study has reported the high prevalence of hypokalemia among patients with COVID-19¹⁸.

178 Potassium derangement is independently associated with increased mortality in ICU patients¹⁹.
179 Deviations in serum potassium levels in COVID-19 patients may suggest dysregulation of the
180 renin-angiotensin system²⁰ which has been suggested to also play a role in SARS-CoV-2
181 pathogenesis. This finding shows that the model aligns with previously reported clinically
182 relevant markers and also predicts new markers that emerged from our patient population.

183 Our study utilizes a multi-hospital cohort that has been developed and validated in
184 temporarily distinct subsets of the cohort. Multiple studies in the past using machine learning
185 methodology to study COVID-19 outcomes used only a few machine learning algorithms^{8-10,21,22}.
186 However, these studies were oriented toward identifying clinical features rather than determining
187 the best machine learning algorithm at predicting clinical outcomes, so only limited number of
188 models were tested. To our knowledge, this is the first study to quantitatively and systematically
189 compare 18 machine learning models through robust methodology encompassing all categories
190 of algorithms. We showed that ensemble-methods perform better than other methods in
191 predicting ICU admission and mortality from COVID-19. Ensemble methods are meta-
192 algorithms that combine several different machine learning techniques into one unified
193 predictive model (Table S3)²³, which could explain this improvement in performance. We have
194 also done exhaustive hyperparameter tuning to determine the best values. By performing SHAP
195 analysis, we showed how variables impact outcomes in black-box machine learning models.
196 Thus, our study is consistent with previous clinical study results, revealing similar clinical
197 predictors for ICU admission and mortality, utilizing higher-performing machine learning
198 models.

199 There are a number of limitations in our study. The lack of complete laboratory values
200 for all patients necessitated exclusion of a large number of patients and removal of some

201 variables in development of the models. As suggested by Jakobsen et al²⁴, imputation is not an
202 advisable method to handle missingness, when the percentage of missing data exceeds 40%. The
203 majority of individuals (>98%) included in our analysis were those patients who visited to ED
204 and subsequently became in-patients. In the patients excluded due to missingness, only ~40% of
205 the patients needed in-patient care. This discrepancy in severity might be the reason for lack of
206 laboratory values in excluded patients.

207 Another limitation is that, as some of the laboratory values may take hours to be reported,
208 the data may not be available until after the patient has transitioned out of the ER, decreasing the
209 utility of using these predictors in triaging patient disposition. Similarly, as the mortality model
210 uses ventilator use as a predictor, it requires ICU admission to be utilized and would not be valid
211 in an earlier phase of care.

212 We also observed that the predicting capability on the external cohort (imbalanced
213 dataset) was higher for ICU admission models in comparison to mortality models. This could be
214 due to the changes instated in the ICU during the later period of pandemic. The changes in the
215 treatment regimens might be affecting the mortality and thereby affecting the predictive power of
216 our models. Our cohort is based on the population from Southern New England region of United
217 States and includes two hospitals that are world-class academic centers, which could also limit
218 the versatility of the models. More elaborate studies based on this framework in other cohorts
219 would help validate our findings.

220 Our model development process and findings could be used by clinicians in gauging the
221 clinical course, particularly ICU admission, of an individual with COVID-19 during an ED
222 encounter. We would recommend using ensemble-based methods for developing clinical
223 prediction models. Our ensemble methods identified key features in patients, such as kidney

224 function, potassium, procalcitonin, CRP and LDH, that allowed us to predict clinical outcomes.
225 Deploying such models would augment the clinical decision-making process by allowing
226 physicians to identify potentially high-risk individuals and adjust their treatment accordingly.

227

228 **Methods**

229 *Study population*

230 Patients from the Mass General Brigham (MGB) healthcare system that were positive for
231 COVID-19 between March and August of 2020 and had an ED encounter were included.
232 Patients either had COVID-19 prior to the index ED visit or were diagnosed during that
233 encounter. MGB is an integrated health care system which encompasses 14 hospitals across the
234 New England area in the United States. COVID-19 positive patients were defined by the
235 COVID-19 infection status, a discretely recorded field in the Epic EHR (Epic, Inc., Verona, WI).
236 The COVID-19 infection status was added automatically if a SARS-CoV-2 PCR test was
237 positive, or by Infection Control personnel if the patient has a confirmed positive test from an
238 outside facility. This study was approved by the MGB Institutional Review Board.

239 *Data collection and covariate selection*

240 We queried the data warehouse of our EHR for patient-level data including
241 demographics, comorbidities, home medications, most recent outpatient recorded blood pressure,
242 and death date. For each hospital encounter we extracted vital signs, laboratory values, admitting
243 service, hospital length of stay, date of first ICU admission, amongst others. The patient's
244 problem list was extracted and transformed into a comorbidity matrix by using the "comorbidity"
245 R package²⁵.

246 ***Outcome definition***

247 The two primary outcomes used for developing the models were ICU admission within 5
248 days of ED encounter and mortality within 28 days of ED encounter. The beginning of the
249 prediction window began upon arrival to the ED.

250 ***Model development***

251 As described in Table S1, we selected a reduced set of potential predictor variables from
252 previously published literature (Table S2). We used the same covariates in developing the ICU
253 admission and mortality models except for ventilator use which was added to mortality models
254 but excluded from ICU admission models. Age (10 year intervals), race (African American or
255 other), BMI, modified Charlson Comorbidity Index ²⁶, angiotensin converting enzyme
256 inhibitor/angiotensin receptor blocker (ACEi/ARB) use, hypertension (>140/90 mmHg), and
257 eGFR <60 ml/min were treated as categorical values. Patients with missing values for the
258 independent variables or obviously incorrect entries (e.g., one patient was listed with respiratory
259 rate of 75 breaths per minute) were excluded. Imputation was not advisable due to a high
260 percentage of missingness²⁴. Models were developed using the patients admitted during the
261 period of March and April 2020. For external validation, we used a temporally distinct cohort
262 consisting of patients admitted from May through August 2020. The data set was imbalanced
263 with significantly fewer patients who were admitted to the ICU or died due to COVID-19
264 compared with those who did not. For the purpose of developing and internally validating the
265 machine learning models, we randomly selected surviving patients who were not admitted to the
266 ICU and matched the number of patients who were admitted to the ICU or died (n = 684 for ICU
267 models and n = 434 for mortality models). From this group of patients, 70% (n = 478 for ICU

268 models and $n = 303$ for mortality models) were used for developing machine learning models
269 and the remaining 30% were used for internal validation.

270 A total of eighteen machine learning algorithms were tested, the descriptions of which are
271 available in Table S3. For every machine learning model, we used a three-step approach. First,
272 we made models using various combinations of tunable hyperparameters which are used to
273 control the learning process of algorithms. The hyperparameters that were adjusted depended on
274 the algorithm (outlined in Table S4). After developing these models for each combination of
275 hyperparameter, we tested the performance of each of these combinations using a cross
276 validation technique (number of folds = 5) during which a precision-recall area under curve (PR
277 AUC) score was considered to select the best hyperparameter (Table S5). PR AUC score
278 compares the positive predictive value (precision) and true positive rate (sensitivity or recall) of a
279 model. For grading the performance of models, we used PR AUC scores as this is more
280 applicable for datasets that are imbalanced. In our case, the external validation dataset remained
281 an imbalanced dataset.

282 ***Evaluation of model performance***

283 Model performance evaluation was done in three parts. A *StratifiedKfold* technique of
284 cross validation was first used during model development. In this method, 20% of the patients
285 were excluded while training the model and the excluded patients were then used to test the
286 model. This was done in an iterative process. Each model was evaluated by calculating the
287 Receiver Operating Characteristic Area Under the Curve (ROC AUC), PR AUC, F1, recall,
288 precision, balanced accuracy, and Brier scores. To calculate the 95% confidence interval, we
289 used $t_{0.975, df=4} = 2.776$ based on t -distribution for $n = 5$.

290 For the second level of validation, the model performance was evaluated on the 30% of
291 patients who were not used during development of the models. This cohort worked as an internal
292 validation dataset for these models. Finally, for the external validation, the cohort of patients
293 who presented to the ED between May and August 2020 was used (Table S6).

294 ***Model interpretation using Shapley values***

295 For explaining the models, SHAP feature importance was reported based on Shapley
296 values²⁷, details of which are outlined in the Supplementary Methods. SHAP values are useful to
297 explain “black-box” machine learning models which are otherwise difficult to interpret. SHAP
298 values for each patient feature explain the intensity and direction of impact on predicting the
299 outcome.

300 ***Software***

301 Data cleaning and processing were performed with R (R Core Team, version 3.6.3) using
302 the tidyverse and comorbidity packages^{25,28,29}. Machine learning model development was done
303 using Python (details in Supplementary Methods)³⁰⁻³³. The programming code for R and Python
304 are available upon request addressed to the corresponding author (jain@steele.mgh.harvard.edu).

305 **Supplementary Materials**

306 Methods

307 Fig. S1. Matrix plots showing differential model performance

308 Fig. S2. ROC AUC and PR AUC plots

309 Table S1. Selection of patients and variable details used for developing and testing the models

310 Table S2. Risk factors identified for mortality and ICU admission in COVID-19 studies

311 Table S3. Description of machine learning algorithms

312 Table S4. Hyperparameters which were optimized for machine learning algorithms

313 Table S5. Best hyperparameter values for machine learning algorithms that were chosen after
314 tuning hyperparameters using *GridSearchCV* and *cross validation* technique.

315 Table S6. Characteristics of patients who visited the emergency room between May and August
316 2020 for COVID-19, that were used to evaluate the machine learning models as an external
317 dataset. Variables stratified based on ICU admission and death of patients.

318 Table S7. Multiple comparison between ensemble methods and other types of machine learning
319 algorithms using Fischer Least Significant Difference (LSD) t-test.

320 **References:**

- 321 1 WorldHealthOrganization. Coronavirus disease (COVID-19): situation report, 182. (2020).
- 322 2 Antommaria, A. H. M. *et al.* Ventilator Triage Policies During the COVID-19 Pandemic at U.S. Hospitals
323 Associated With Members of the Association of Bioethics Program Directors. *Ann Intern Med* **173**, 188-
324 194, doi:10.7326/m20-1738 (2020).
- 325 3 Silberzweig, J. *et al.* Rationing Scarce Resources: The Potential Impact of COVID-19 on Patients with
326 Chronic Kidney Disease. *Journal of the American Society of Nephrology* **31**, 1926,
327 doi:10.1681/ASN.2020050704 (2020).

- 328 4 Tomasev, N. *et al.* A clinically applicable approach to continuous prediction of future acute kidney injury.
329 *Nature* **572**, 116-119, doi:10.1038/s41586-019-1390-1 (2019).
- 330 5 Henry, K. E., Hager, D. N., Pronovost, P. J. & Saria, S. A targeted real-time early warning score
331 (TREWScore) for septic shock. *Sci Transl Med* **7**, 299ra122, doi:10.1126/scitranslmed.aab3719 (2015).
- 332 6 Wehbe, R. M., Khan, S. S., Shah, S. J. & Ahmad, F. S. Predicting High-Risk Patients and High-Risk
333 Outcomes in Heart Failure. *Heart Fail Clin* **16**, 387-407, doi:10.1016/j.hfc.2020.05.002 (2020).
- 334 7 Subudhi, S., Verma, A. & B.Patel, A. Prognostic machine learning models for COVID-19 to facilitate
335 decision making. *International Journal of Clinical Practice*, e13685, doi:10.1111/ijcp.13685 (2020).
- 336 8 Yan, L. *et al.* An interpretable mortality prediction model for COVID-19 patients. *Nature Machine*
337 *Intelligence* **2**, 283-288, doi:10.1038/s42256-020-0180-7 (2020).
- 338 9 Iwendi, C. *et al.* COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm. *Front*
339 *Public Health* **8**, 357, doi:10.3389/fpubh.2020.00357 (2020).
- 340 10 Wu, G. *et al.* Development of a clinical decision support system for severity risk prediction and triage of
341 COVID-19 patients at hospital admission: an international multicentre study. *Eur Respir J* **56**,
342 doi:10.1183/13993003.01104-2020 (2020).
- 343 11 Eaneff, S., Obermeyer, Z. & Butte, A. J. The Case for Algorithmic Stewardship for Artificial Intelligence
344 and Machine Learning Technologies. *JAMA*, doi:10.1001/jama.2020.9371 (2020).
- 345 12 Lippi, G. & Plebani, M. Procalcitonin in patients with severe coronavirus disease 2019 (COVID-19): A
346 meta-analysis. *Clin Chim Acta* **505**, 190-191, doi:10.1016/j.cca.2020.03.004 (2020).
- 347 13 Linscheid, P. *et al.* In vitro and in vivo calcitonin I gene expression in parenchymal cells: a novel product
348 of human adipose tissue. *Endocrinology* **144**, 5578-5584, doi:10.1210/en.2003-0854 (2003).
- 349 14 Muller, B. *et al.* Ubiquitous expression of the calcitonin-i gene in multiple tissues in response to sepsis. *J*
350 *Clin Endocrinol Metab* **86**, 396-404, doi:10.1210/jcem.86.1.7089 (2001).
- 351 15 Meisner, M. Update on procalcitonin measurements. *Ann Lab Med* **34**, 263-273,
352 doi:10.3343/alm.2014.34.4.263 (2014).
- 353 16 Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**,
354 430-436, doi:10.1038/s41586-020-2521-4 (2020).

- 355 17 Flythe, J. E. *et al.* Characteristics and Outcomes of Individuals With Pre-existing Kidney Disease and
356 COVID-19 Admitted to Intensive Care Units in the United States. *Am J Kidney Dis*,
357 doi:10.1053/j.ajkd.2020.09.003 (2020).
- 358 18 Chen, D. *et al.* Assessment of Hypokalemia and Clinical Characteristics in Patients With Coronavirus
359 Disease 2019 in Wenzhou, China. *JAMA Netw Open* **3**, e2011122,
360 doi:10.1001/jamanetworkopen.2020.11122 (2020).
- 361 19 Hessels, L. *et al.* The relationship between serum potassium, potassium variability and in-hospital mortality
362 in critically ill patients and a before-after analysis on the impact of computer-assisted potassium control.
363 *Crit Care* **19**, 4, doi:10.1186/s13054-014-0720-9 (2015).
- 364 20 Weir, M. R. & Rolfe, M. Potassium homeostasis and renin-angiotensin-aldosterone system inhibitors. *Clin*
365 *J Am Soc Nephrol* **5**, 531-548, doi:10.2215/CJN.07821109 (2010).
- 366 21 Yao, H. *et al.* Severity Detection for the Coronavirus Disease 2019 (COVID-19) Patients Using a Machine
367 Learning Model Based on the Blood and Urine Tests. *Front Cell Dev Biol* **8**, 683,
368 doi:10.3389/fcell.2020.00683 (2020).
- 369 22 Liang, W. *et al.* Early triage of critically ill COVID-19 patients using deep learning. *Nat Commun* **11**, 3543,
370 doi:10.1038/s41467-020-17280-8 (2020).
- 371 23 Dietterich, T. G. in *Proceedings of the First International Workshop on Multiple Classifier Systems* 1–15
372 (Springer-Verlag, 2000).
- 373 24 Jakobsen, J. C., Gluud, C., Wetterslev, J. & Winkel, P. When and how should multiple imputation be used
374 for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Med Res*
375 *Methodol* **17**, 162, doi:10.1186/s12874-017-0442-1 (2017).
- 376 25 Gasparini, A. comorbidity: An R package for computing comorbidity scores. *Journal of Open Source*
377 *Software* **3**, 648, doi:10.21105/joss.00648 (2018).
- 378 26 Charlson, M. E., Pompei, P., Ales, K. L. & MacKenzie, C. R. A new method of classifying prognostic
379 comorbidity in longitudinal studies: development and validation. *J Chronic Dis* **40**, 373-383,
380 doi:10.1016/0021-9681(87)90171-8 (1987).
- 381 27 Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. 4765--4774 (2017).

- 382 28 R_Core_Team. *R: A Language and Environment for Statistical Computing*, <<https://www.R-project.org/>>
383 (2020).
- 384 29 Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686,
385 doi:10.21105/joss.01686 (2019).
- 386 30 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**
387 (2012).
- 388 31 Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357-362, doi:10.1038/s41586-020-2649-
389 2 (2020).
- 390 32 Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System*. (2016).
- 391 33 McKinney, W. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python High*
392 *Performance Science Computer* (2011).

393

394

395 **Acknowledgments:** Rakesh Jain's research is supported by R01-CA208205, and U01-CA
396 224348, Outstanding Investigator Award R35-CA197743 and grants from the National
397 Foundation for Cancer Research, Jane's Trust Foundation, American Medical Research
398 Foundation and Harvard Ludwig Cancer Center. We would like to thank Ashwin
399 Srinivasan Kumar, Avanish Ranjan, Tariq Anwar and Mushtaq Rizvi for advice on
400 machine learning algorithms and Python coding. **Author contributions:** S.S. performed,
401 designed and built machine learning models. S.D. extracted data from the MGB database.
402 A.V., A.B.P., C.C.H., M.J.K., S.D., H.L., T.S., L.L.M., and R.K.J supervised model
403 development. All authors were involved in writing the article. **Competing interests:**
404 LLM owns equity in Bayer AG and is a consultant for SimBiosys. R.K.J. received
405 honorarium from Amgen; consultant fees from Chugai, Merck, Ophthotech, Pfizer,
406 SPARC, SynDevRx, XTuit; owns equity in Accurius, Enlight, Ophthotech, SynDevRx;

407 and serves on the Boards of Trustees of Tekla Healthcare Investors, Tekla Life Sciences
408 Investors, Tekla Healthcare Opportunities Fund, Tekla World Healthcare Fund. Neither
409 any reagent nor any funding from these organizations was used in this study. Other
410 coauthors have no conflict of interests to declare. **Data and materials availability:** The
411 programming code for R and Python are available upon request addressed to the
412 corresponding author (jain@steele.mgh.harvard.edu).

413
414

415 **Figures legends:**

416

417 **Fig. 1.** Schematic diagram representing the process of machine learning model development. **(A)**
418 Flow diagram depicting steps in obtaining the training and external validation datasets
419 (with patient numbers in each step). **(B)** The process of patient selection, dataset
420 balancing, hyperparameter tuning, cross-validation, internal and external validation are
421 shown.

422

423 **Fig. 2. (A-B). Boxplots** representing the precision recall area under the curve (PR AUC) and
424 receiver operating characteristic area under the curve (ROC AUC) scores of ICU
425 admission and mortality prediction models. Error bars indicate minimum and maximum
426 values. Statistical analysis was performed using Fisher's Least Significant Difference
427 (LSD) *t*-test. *p*-value style is geometric progression - <0.03 (*), <0.002 (**), <0.0002
428 (***), <0.0001 (****). **Variables of importance for ICU admission and mortality
429 prediction models.** (C) SHAP value summary dot plot and (D) variable of importance of
430 *RandomForest* algorithm-based ICU admission model. (E) SHAP value summary dot

431 plot and (F) variable of importance of *XGBClassifier* algorithm-based mortality model.
432 The calculation of SHAP values is done by comparing the prediction of the model with
433 and without the feature in every possible way of adding the feature to the model. The bar
434 plot depicts the mean SHAP values whereas the summary dot plot shows the impact on
435 the model. The color of the dot represents the value of the feature and the X-axis depicts
436 the direction and magnitude of the impact. Red colored dots represent high value of the
437 feature and the blue represents lower value. A positive SHAP value means the feature
438 value increases likelihood of ICU admission/mortality. For features with positive SHAP
439 value for red dots, suggests directly proportional variable to outcome of interest and those
440 with positive SHAP value for blue dots, suggest inverse correlation.

441 **Table 1.** Characteristics of patients who visited emergency department during March and April 2020 for COVID-19, that were
 442 included for building the machine learning models. Variables stratified based on ICU admission and death of patients.

n	ICU admission			p	Death		
	Overall 1144	No 802	Yes 342		No 927	Yes 217	p
Demographics							
Age group (%)				<0.001			<0.001
10-19	1 (0.1)	1 (0.1)	0 (0.0)	0	1 (0.1)	0 (0.0)	0
20-29	26 (2.3)	19 (2.4)	7 (2.0)	0	26 (2.8)	0 (0.0)	0
30-39	71 (6.2)	53 (6.6)	18 (5.3)	0	70 (7.6)	1 (0.5)	0
40-49	113 (9.9)	80 (10.0)	33 (9.6)	0	108 (11.7)	5 (2.3)	0
50-59	196 (17.1)	131 (16.3)	65 (19.0)	0	180 (19.4)	16 (7.4)	0
60-69	223 (19.5)	131 (16.3)	92 (26.9)	0	200 (21.6)	23 (10.6)	0
70-79	231 (20.2)	158 (19.7)	73 (21.3)	0	169 (18.2)	62 (28.6)	0
80-89	201 (17.6)	152 (19.0)	49 (14.3)	0	126 (13.6)	75 (34.6)	0
90+	82 (7.2)	77 (9.6)	5 (1.5)	0	47 (5.1)	35 (16.1)	0
Sex = Male (%)	629 (55.0)	402 (50.1)	227 (66.4)	<0.001	502 (54.2)	127 (58.5)	0.276
Race = Other (%)	949 (83.0)	675 (84.2)	274 (80.1)	0.114	772 (83.3)	177 (81.6)	0.615
BMI_categorical (%)				0.005			0.001
[0,25]	285 (24.9)	220 (27.4)	65 (19.0)	0	209 (22.5)	76 (35.0)	0
(25,30]	387 (33.8)	270 (33.7)	117 (34.2)	0	326 (35.2)	61 (28.1)	0
(30,75]	472 (41.3)	312 (38.9)	160 (46.8)	0	392 (42.3)	80 (36.9)	0
Medication use							
On ACEi/ARB = TRUE (%)	288 (25.2)	190 (23.7)	98 (28.7)	0.09	232 (25.0)	56 (25.8)	0.88
On ARA = TRUE (%)	26 (2.3)	16 (2.0)	10 (2.9)	0.454	17 (1.8)	9 (4.1)	0.071
On Calcium channel blocker = TRUE (%)	220 (19.2)	157 (19.6)	63 (18.4)	0.71	163 (17.6)	57 (26.3)	0.005
On Betablocker = TRUE (%)	285 (24.9)	208 (25.9)	77 (22.5)	0.25	195 (21.0)	90 (41.5)	<0.001
On Vasodilator = TRUE (%)	80 (7.0)	64 (8.0)	16 (4.7)	0.06	54 (5.8)	26 (12.0)	0.002
On Alphablocker = TRUE (%)	19 (1.7)	16 (2.0)	3 (0.9)	0.271	16 (1.7)	3 (1.4)	0.951
On Diuretic = TRUE (%)	250 (21.9)	187 (23.3)	63 (18.4)	0.079	174 (18.8)	76 (35.0)	<0.001
On Antiplatelet = TRUE (%)	35 (3.1)	26 (3.2)	9 (2.6)	0.718	28 (3.0)	7 (3.2)	1
On NSAIDs = TRUE (%)	126 (11.0)	84 (10.5)	42 (12.3)	0.429	113 (12.2)	13 (6.0)	0.012
On Proton pump inhibitor = TRUE (%)	275 (24.0)	191 (23.8)	84 (24.6)	0.846	206 (22.2)	69 (31.8)	0.004
On Statin = TRUE (%)	456 (39.9)	326 (40.6)	130 (38.0)	0.443	332 (35.8)	124 (57.1)	<0.001
On Anticoagulant = TRUE (%)	133 (11.6)	93 (11.6)	40 (11.7)	1	84 (9.1)	49 (22.6)	<0.001
History of past illness							
Acute myocardial infarction = 1 (%)	33 (2.9)	25 (3.1)	8 (2.3)	0.598	21 (2.3)	12 (5.5)	0.018
Congestive heart failure = 1 (%)	136 (11.9)	107 (13.3)	29 (8.5)	0.026	79 (8.5)	57 (26.3)	<0.001
Peripheral vascular disease = 1 (%)	78 (6.8)	58 (7.2)	20 (5.8)	0.47	51 (5.5)	27 (12.4)	<0.001
Cerebrovascular disease = 1 (%)	109 (9.5)	81 (10.1)	28 (8.2)	0.369	65 (7.0)	44 (20.3)	<0.001
Dementia = 1 (%)	78 (6.8)	67 (8.4)	11 (3.2)	0.002	43 (4.6)	35 (16.1)	<0.001
Chronic obstructive pulmonary disease = 1 (%)	167 (14.6)	132 (16.5)	35 (10.2)	0.008	118 (12.7)	49 (22.6)	<0.001
Rheumatic disease = 1 (%)	31 (2.7)	21 (2.6)	10 (2.9)	0.926	22 (2.4)	9 (4.1)	0.224
Peptic ulcer disease = 1 (%)	16 (1.4)	13 (1.6)	3 (0.9)	0.48	11 (1.2)	5 (2.3)	0.347
Mild liver disease = 1 (%)	70 (6.1)	49 (6.1)	21 (6.1)	1	54 (5.8)	16 (7.4)	0.484

Diabetes = 1 (%)	257 (22.5)	169 (21.1)	88 (25.7)	0.099	189 (20.4)	68 (31.3)	0.001
Diabetes with complications = 1 (%)	113 (9.9)	81 (10.1)	32 (9.4)	0.781	77 (8.3)	36 (16.6)	<0.001
Hemiplegia = 1 (%)	12 (1.0)	7 (0.9)	5 (1.5)	0.563	9 (1.0)	3 (1.4)	0.868
Renal disease = 1 (%)	180 (15.7)	136 (17.0)	44 (12.9)	0.099	104 (11.2)	76 (35.0)	<0.001
Cancer = 1 (%)	133 (11.6)	100 (12.5)	33 (9.6)	0.207	89 (9.6)	44 (20.3)	<0.001
Moderate/severe liver disease = 1 (%)	9 (0.8)	5 (0.6)	4 (1.2)	0.554	6 (0.6)	3 (1.4)	0.499
Metastatic cancer = 1 (%)	14 (1.2)	11 (1.4)	3 (0.9)	0.687	7 (0.8)	7 (3.2)	0.008
AIDS = 1 (%)	9 (0.8)	8 (1.0)	1 (0.3)	0.384	6 (0.6)	3 (1.4)	0.499
Hypertension = 1 (%)	464 (40.6)	318 (39.7)	146 (42.7)	0.372	372 (40.1)	92 (42.4)	0.592
Laboratory values and clinical examination							
CRP (mg/L) (mean (SD))	97.74 (82.36)	76.50 (68.54)	147.54 (90.31)	<0.001	91.98 (80.07)	122.35 (87.53)	<0.001
First respiratory rate (counts/min) (mean (SD))	24.23 (7.19)	22.78 (5.85)	27.63 (8.74)	<0.001	23.71 (6.67)	26.45 (8.76)	<0.001
First heart rate (beats/min) (mean (SD))	95.44 (19.67)	92.89 (18.87)	101.41 (20.22)	<0.001	96.03 (19.24)	92.89 (21.28)	0.034
Sodium (mmol/L) (mean (SD))	137.36 (5.56)	137.68 (5.09)	136.61 (6.48)	0.003	136.99 (5.05)	138.94 (7.14)	<0.001
Calcium (mg/dL) (mean (SD))	8.98 (0.59)	9.03 (0.59)	8.86 (0.59)	<0.001	8.98 (0.58)	8.95 (0.64)	0.46
Magnesium (mg/dL) (mean (SD))	2.03 (0.33)	2.01 (0.31)	2.08 (0.38)	0.003	2.01 (0.32)	2.11 (0.36)	<0.001
Potassium (mmol/L) (mean (SD))	4.11 (0.59)	4.10 (0.57)	4.12 (0.65)	0.604	4.04 (0.54)	4.37 (0.72)	<0.001
Chloride (mmol/L) (mean (SD))	98.46 (5.95)	98.88 (5.50)	97.48 (6.78)	<0.001	98.09 (5.55)	100.03 (7.22)	<0.001
Lymphocytes (percentage; ref = 22-44%) (mean (SD))	16.60 (10.17)	17.87 (10.20)	13.61 (9.47)	<0.001	17.20 (9.67)	14.02 (11.76)	<0.001
Neutrophils (percentage; ref = 40-70%) (mean (SD))	73.30 (12.14)	71.65 (12.16)	77.17 (11.20)	<0.001	72.80 (11.60)	75.43 (14.04)	0.004
WBC (x1000/ μ L) (mean (SD))	7.61 (6.15)	7.19 (5.63)	8.60 (7.15)	<0.001	7.41 (5.43)	8.48 (8.55)	0.02
D-dimer (ng/mL) (mean (SD))	1923.26 (3473.28)	1749.62 (2383.80)	2330.43 (5181.81)	0.01	1779.02 (3526.78)	2539.41 (3169.14)	0.004
Total bilirubin (mg/dL) (mean (SD))	0.58 (0.89)	0.56 (1.02)	0.62 (0.43)	0.322	0.57 (0.96)	0.62 (0.49)	0.45
Ferritin (μ g/L) (mean (SD))	935.85 (2071.69)	738.57 (1066.99)	1398.46 (3377.22)	<0.001	829.81 (1167.77)	1388.82 (4075.61)	<0.001
LDH (Units) (mean (SD))	370.12 (517.90)	321.17 (273.25)	484.93 (839.54)	<0.001	348.47 (271.12)	462.65 (1045.77)	0.003
Low GFR (<60 ml/min/1.73m ²) = TRUE (%)	454 (39.7)	317 (39.5)	137 (40.1)	0.918	300 (32.4)	154 (71.0)	<0.001
Anion gap (mmol/L) (mean (SD))	15.71 (3.49)	15.24 (3.16)	16.81 (3.96)	<0.001	15.52 (3.31)	16.54 (4.08)	<0.001
Hemoglobin (g/dL) (mean (SD))	12.99 (2.09)	12.84 (2.06)	13.32 (2.12)	<0.001	13.10 (1.98)	12.52 (2.47)	<0.001
First O ₂ saturation (%) (mean (SD))	93.81 (6.22)	94.93 (4.66)	91.17 (8.30)	<0.001	94.06 (6.16)	92.71 (6.40)	0.004
ventilator_use = TRUE (%)	294 (25.7)	NA	NA	NA	195 (21.0)	99 (45.6)	<0.001
Procalcitonin (ng/ml) (mean (SD))	1.11 (5.98)	0.63 (3.56)	2.23 (9.40)	<0.001	0.73 (3.99)	2.73 (10.86)	<0.001
Glucose (mg/dL) (mean (SD))	153.76 (80.04)	146.45 (70.68)	170.91 (96.53)	<0.001	151.32 (78.94)	164.19 (83.96)	0.033
ALT (IU/L) (mean (SD))	45.05 (202.49)	35.68 (36.49)	67.02 (365.53)	0.016	40.73 (39.82)	63.50 (457.98)	0.136

443 **Table 2.** Performance of machine learning models to predict ICU admission within 5 days in COVID-19 patients. Cross-validation
 444 scores are expressed as mean \pm 95% confidence interval.

Method Type	Model Name	Dataset	ROC AUC	PR AUC	F1 score	Recall	Precision	Balanced	Brier	Total
-------------	------------	---------	---------	--------	----------	--------	-----------	----------	-------	-------

			accuracy	score	positive events						
Ensemble	AdaBoostClassifier	Cross-validation	0.79 ± 0.06	0.8 ± 0.07	0.74 ± 0.04	0.74 ± 0.03	0.74 ± 0.06	0.73 ± 0.05	0.23 ± 0.0	244/478	
		Internal validation	0.8	0.81	0.67	0.63	0.72	0.71	0.23	98/206	
		External validation	0.76	0.54	0.53	0.62	0.46	0.71	0.23	74/334	
	BaggingClassifier	Cross-validation	0.8 ± 0.04	0.8 ± 0.07	0.73 ± 0.06	0.71 ± 0.07	0.76 ± 0.06	0.73 ± 0.05	0.2 ± 0.01	244/478	
		Internal validation	0.81	0.81	0.73	0.69	0.77	0.75	0.2	98/206	
		External validation	0.79	0.61	0.54	0.62	0.48	0.72	0.18	74/334	
	GradientBoostingClassifier	Cross-validation	0.77 ± 0.07	0.77 ± 0.09	0.73 ± 0.07	0.73 ± 0.1	0.73 ± 0.06	0.72 ± 0.06	0.2 ± 0.04	244/478	
		Internal validation	0.8	0.8	0.7	0.66	0.73	0.72	0.19	98/206	
		External validation	0.77	0.57	0.52	0.66	0.43	0.7	0.19	74/334	
	RandomForestClassifier	Cross-validation	0.79 ± 0.06	0.8 ± 0.1	0.74 ± 0.06	0.73 ± 0.06	0.76 ± 0.07	0.74 ± 0.07	0.19 ± 0.02	244/478	
		Internal validation	0.8	0.82	0.72	0.67	0.77	0.74	0.18	98/206	
		External validation	0.81	0.62	0.56	0.66	0.49	0.73	0.16	74/334	
	XGBClassifier	Cross-validation	0.78 ± 0.06	0.78 ± 0.08	0.73 ± 0.04	0.72 ± 0.05	0.74 ± 0.04	0.72 ± 0.04	0.2 ± 0.03	244/478	
		Internal validation	0.81	0.81	0.7	0.66	0.74	0.73	0.18	98/206	
		External validation	0.77	0.6	0.51	0.59	0.45	0.69	0.17	74/334	
ExtraTreesClassifier	Cross-validation	0.79 ± 0.04	0.79 ± 0.07	0.72 ± 0.06	0.71 ± 0.08	0.73 ± 0.04	0.72 ± 0.05	0.19 ± 0.01	244/478		
	Internal validation	0.79	0.8	0.67	0.65	0.7	0.7	0.19	98/206		
	External validation	0.75	0.54	0.49	0.64	0.39	0.68	0.19	74/334		
Gaussian process	GaussianProcessClassifier	Cross-validation	0.63 ± 0.06	0.6 ± 0.06	0.55 ± 0.07	0.48 ± 0.05	0.64 ± 0.1	0.59 ± 0.07	0.25 ± 0.0	244/478	
		Internal validation	0.58	0.5	0.48	0.45	0.52	0.54	0.25	98/206	
		External validation	0.65	0.29	0.31	0.34	0.29	0.55	0.25	74/334	
Linear models	LogisticRegression	Cross-validation	0.77 ± 0.07	0.79 ± 0.08	0.71 ± 0.04	0.69 ± 0.05	0.73 ± 0.05	0.71 ± 0.05	0.19 ± 0.03	244/478	
		Internal validation	0.83	0.83	0.73	0.7	0.77	0.75	0.17	98/206	
		External validation	0.81	0.62	0.58	0.64	0.53	0.74	0.16	74/334	
	PassiveAggressiveClassifier	Cross-validation	0.67 ± 0.1	0.7 ± 0.09	0.49 ± 0.32	0.59 ± 0.48	0.69 ± 0.3	0.53 ± 0.09	0.28 ± 0.06	244/478	
		Internal validation	0.77	0.77	0.1	0.05	1	0.53	0.34	98/206	
		External validation	0.73	0.46	0.17	0.09	0.78	0.54	0.16	74/334	
	SGDClassifier	Cross-validation	0.68 ± 0.03	0.63 ± 0.03	0.69 ± 0.04	0.69 ± 0.05	0.69 ± 0.03	0.68 ± 0.03	0.32 ± 0.03	244/478	
		Internal validation	0.72	0.65	0.69	0.63	0.76	0.72	0.27	98/206	
		External validation	0.7	0.39	0.53	0.54	0.53	0.7	0.21	74/334	
	Perceptron	Cross-validation	0.71 ± 0.06	0.72 ± 0.05	0.37 ± 0.39	0.39 ± 0.54	0.78 ± 0.22	0.57 ± 0.1	0.32 ± 0.11	244/478	
		Internal validation	0.71	0.72	0.64	0.99	0.47	0.49	0.31	98/206	
		External validation	0.58	0.35	0.36	0.97	0.22	0.49	0.57	74/334	
	Naïve Bayes	GaussianNB	Cross-validation	0.72 ± 0.03	0.71 ± 0.08	0.57 ± 0.09	0.47 ± 0.12	0.74 ± 0.06	0.65 ± 0.04	0.34 ± 0.03	244/478
			Internal validation	0.75	0.74	0.58	0.46	0.8	0.68	0.3	98/206
			External validation	0.71	0.46	0.48	0.5	0.46	0.67	0.22	74/334
Nearest Neighbor	KNeighborsClassifier	Cross-validation	0.67 ± 0.06	0.68 ± 0.07	0.62 ± 0.04	0.58 ± 0.05	0.66 ± 0.06	0.63 ± 0.05	0.23 ± 0.01	244/478	
		Internal validation	0.71	0.7	0.66	0.68	0.64	0.67	0.22	98/206	
		External validation	0.69	0.45	0.44	0.59	0.35	0.64	0.2	74/334	
Support vector machine	LinearSVC	Cross-validation	0.59 ± 0.1	0.65 ± 0.08	0.39 ± 0.36	0.46 ± 0.56	0.67 ± 0.29	0.52 ± 0.08	0.33 ± 0.07	244/478	
		Internal validation	0.63	0.58	0.02	0.01	1	0.51	0.44	98/206	
		External validation	0.67	0.38	0.05	0.03	1	0.51	0.21	74/334	
Tree based	DecisionTreeClassifier	Cross-validation	0.66 ± 0.08	0.67 ± 0.09	0.62 ± 0.1	0.61 ± 0.12	0.63 ± 0.08	0.62 ± 0.08	0.27 ± 0.04	244/478	
		Internal validation	0.69	0.64	0.57	0.53	0.6	0.61	0.25	98/206	
		External validation	0.68	0.4	0.42	0.55	0.34	0.63	0.22	74/334	
Discriminant analysis	LinearDiscriminantAnalysis	Cross-validation	0.74 ± 0.05	0.76 ± 0.08	0.69 ± 0.05	0.68 ± 0.07	0.71 ± 0.04	0.69 ± 0.04	0.22 ± 0.03	244/478	
		Internal validation	0.74	0.74	0.65	0.67	0.63	0.66	0.22	98/206	
		External validation	0.71	0.5	0.46	0.57	0.39	0.66	0.2	74/334	

Neural network	QuadraticDiscriminantAnalysis	Cross-validation	0.72 ± 0.03	0.69 ± 0.03	0.74 ± 0.04	0.87 ± 0.06	0.64 ± 0.03	0.68 ± 0.04	0.31 ± 0.04	244/478
		Internal validation	0.79	0.74	0.71	0.86	0.61	0.68	0.31	98/206
		External validation	0.79	0.48	0.48	0.88	0.33	0.68	0.41	74/334
	MLPClassifier	Cross-validation	0.72 ± 0.05	0.72 ± 0.05	0.7 ± 0.05	0.78 ± 0.16	0.65 ± 0.1	0.65 ± 0.11	0.22 ± 0.03	244/478
		Internal validation	0.77	0.78	0.68	0.63	0.73	0.71	0.19	98/206
		External validation	0.75	0.58	0.53	0.65	0.44	0.71	0.18	74/334

445 **Table 3.** Performance of machine learning models to predict mortality within 28 days in COVID-19 patients. Cross-validation scores
 446 are expressed as mean ± 95% confidence interval.

Method Type	Model Name	Dataset	ROC AUC	PR AUC	F1 score	Recall	Precision	Balanced accuracy	Brier score	Total positive events
Ensemble	AdaBoostClassifier	Cross-validation	0.82 ± 0.05	0.81 ± 0.05	0.73 ± 0.06	0.73 ± 0.09	0.73 ± 0.07	0.73 ± 0.06	0.23 ± 0.0	154/303
		Internal validation	0.79	0.75	0.69	0.68	0.7	0.71	0.24	63/131
		External validation	0.78	0.38	0.42	0.71	0.3	0.73	0.23	45/334
	BaggingClassifier	Cross-validation	0.82 ± 0.03	0.81 ± 0.07	0.75 ± 0.04	0.77 ± 0.06	0.74 ± 0.05	0.74 ± 0.04	0.18 ± 0.01	154/303
		Internal validation	0.78	0.71	0.71	0.78	0.64	0.69	0.19	63/131
		External validation	0.81	0.4	0.4	0.73	0.28	0.72	0.18	45/334
	GradientBoostingClassifier	Cross-validation	0.83 ± 0.07	0.81 ± 0.08	0.76 ± 0.08	0.75 ± 0.06	0.77 ± 0.12	0.75 ± 0.09	0.2 ± 0.06	154/303
		Internal validation	0.78	0.72	0.63	0.6	0.67	0.66	0.27	63/131
		External validation	0.76	0.33	0.35	0.58	0.25	0.66	0.24	45/334
	RandomForestClassifier	Cross-validation	0.81 ± 0.02	0.8 ± 0.05	0.74 ± 0.04	0.77 ± 0.07	0.72 ± 0.05	0.73 ± 0.04	0.2 ± 0.0	154/303
		Internal validation	0.78	0.75	0.71	0.73	0.69	0.71	0.21	63/131
		External validation	0.8	0.37	0.4	0.76	0.27	0.72	0.21	45/334
	XGBClassifier	Cross-validation	0.82 ± 0.05	0.82 ± 0.07	0.74 ± 0.07	0.73 ± 0.1	0.75 ± 0.05	0.74 ± 0.06	0.17 ± 0.03	154/303
		Internal validation	0.79	0.73	0.69	0.7	0.68	0.69	0.19	63/131
		External validation	0.78	0.35	0.42	0.69	0.3	0.72	0.17	45/334
ExtraTreesClassifier	Cross-validation	0.84 ± 0.04	0.82 ± 0.08	0.78 ± 0.01	0.81 ± 0.05	0.76 ± 0.06	0.77 ± 0.02	0.18 ± 0.01	154/303	
	Internal validation	0.77	0.74	0.68	0.71	0.65	0.68	0.2	63/131	
	External validation	0.77	0.31	0.36	0.67	0.25	0.68	0.19	45/334	
Gaussian process	GaussianProcessClassifier	Cross-validation	0.53 ± 0.1	0.55 ± 0.13	0.4 ± 0.09	0.34 ± 0.09	0.49 ± 0.11	0.49 ± 0.07	0.25 ± 0.0	154/303
		Internal validation	0.6	0.54	0.47	0.43	0.53	0.54	0.25	63/131
		External validation	0.54	0.14	0.2	0.38	0.14	0.51	0.25	45/334
Linear models	LogisticRegression	Cross-validation	0.72 ± 0.11	0.73 ± 0.11	0.7 ± 0.1	0.72 ± 0.1	0.68 ± 0.12	0.68 ± 0.11	0.22 ± 0.04	154/303
		Internal validation	0.65	0.65	0.61	0.68	0.56	0.59	0.24	63/131
		External validation	0.66	0.23	0.33	0.58	0.23	0.64	0.21	45/334
	PassiveAggressiveClassifier	Cross-validation	0.71 ± 0.11	0.69 ± 0.12	0.29 ± 0.4	0.27 ± 0.41	0.53 ± 0.52	0.56 ± 0.13	0.31 ± 0.12	154/303
		Internal validation	0.65	0.59	0.64	0.98	0.48	0.49	0.45	63/131
		External validation	0.71	0.32	0.08	0.04	0.67	0.52	0.11	45/334
	SGDClassifier	Cross-validation	0.56 ± 0.09	0.54 ± 0.05	0.4 ± 0.46	0.48 ± 0.55	0.35 ± 0.4	0.56 ± 0.09	0.44 ± 0.09	154/303
		Internal validation	0.53	0.5	0.62	0.83	0.5	0.53	0.48	63/131
		External validation	0.54	0.15	0.24	0.58	0.15	0.54	0.48	45/334
Perceptron	Cross-validation	0.55 ± 0.05	0.6 ± 0.07	0.39 ± 0.35	0.47 ± 0.55	0.54 ± 0.09	0.5 ± 0.03	0.33 ± 0.07	154/303	
	Internal validation	0.55	0.62	0.03	0.02	1	0.51	0.45	63/131	
	External validation	0.55	0.16	0.21	0.38	0.15	0.52	0.26	45/334	
Naïve	GaussianNB	Cross-validation	0.79 ± 0.06	0.77 ± 0.07	0.73 ± 0.04	0.7 ± 0.03	0.76 ± 0.05	0.73 ± 0.04	0.25 ± 0.03	154/303

Bayes		Internal validation	0.75	0.72	0.69	0.71	0.66	0.69	0.28	63/131
		External validation	0.73	0.25	0.35	0.6	0.25	0.66	0.27	45/334
Nearest Neighbor	KNeighborsClassifier	Cross-validation	0.59 ± 0.09	0.6 ± 0.08	0.59 ± 0.04	0.58 ± 0.07	0.6 ± 0.07	0.59 ± 0.05	0.25 ± 0.02	154/303
		Internal validation	0.68	0.66	0.63	0.63	0.62	0.64	0.22	63/131
		External validation	0.61	0.21	0.27	0.53	0.18	0.58	0.23	45/334
Support vector machine	LinearSVC	Cross-validation	0.69 ± 0.11	0.73 ± 0.1	0.66 ± 0.05	0.83 ± 0.24	0.57 ± 0.11	0.57 ± 0.07	0.29 ± 0.11	154/303
		Internal validation	0.62	0.61	0.64	0.98	0.48	0.49	0.47	63/131
		External validation	0.7	0.34	0.09	0.04	1	0.52	0.11	45/334
Tree based	DecisionTreeClassifier	Cross-validation	0.75 ± 0.08	0.72 ± 0.08	0.67 ± 0.04	0.63 ± 0.06	0.72 ± 0.09	0.68 ± 0.05	0.22 ± 0.04	154/303
		Internal validation	0.72	0.67	0.64	0.62	0.67	0.67	0.24	63/131
		External validation	0.7	0.22	0.34	0.58	0.24	0.65	0.25	45/334
Discriminant analysis	LinearDiscriminantAnalysis	Cross-validation	0.85 ± 0.05	0.85 ± 0.06	0.77 ± 0.04	0.79 ± 0.05	0.75 ± 0.06	0.76 ± 0.05	0.18 ± 0.04	154/303
		Internal validation	0.74	0.72	0.65	0.65	0.65	0.66	0.26	63/131
		External validation	0.81	0.34	0.45	0.78	0.32	0.76	0.2	45/334
	QuadraticDiscriminantAnalysis	Cross-validation	0.76 ± 0.08	0.74 ± 0.11	0.72 ± 0.07	0.71 ± 0.13	0.74 ± 0.05	0.72 ± 0.05	0.26 ± 0.04	154/303
		Internal validation	0.77	0.76	0.71	0.76	0.66	0.7	0.27	63/131
		External validation	0.7	0.23	0.35	0.67	0.24	0.67	0.3	45/334
Neural network	MLPClassifier	Cross-validation	0.72 ± 0.11	0.72 ± 0.14	0.71 ± 0.06	0.82 ± 0.08	0.64 ± 0.09	0.66 ± 0.09	0.29 ± 0.09	154/303
		Internal validation	0.72	0.72	0.68	0.94	0.54	0.59	0.38	63/131
		External validation	0.71	0.29	0.3	0.87	0.18	0.63	0.46	45/334

Figure 1

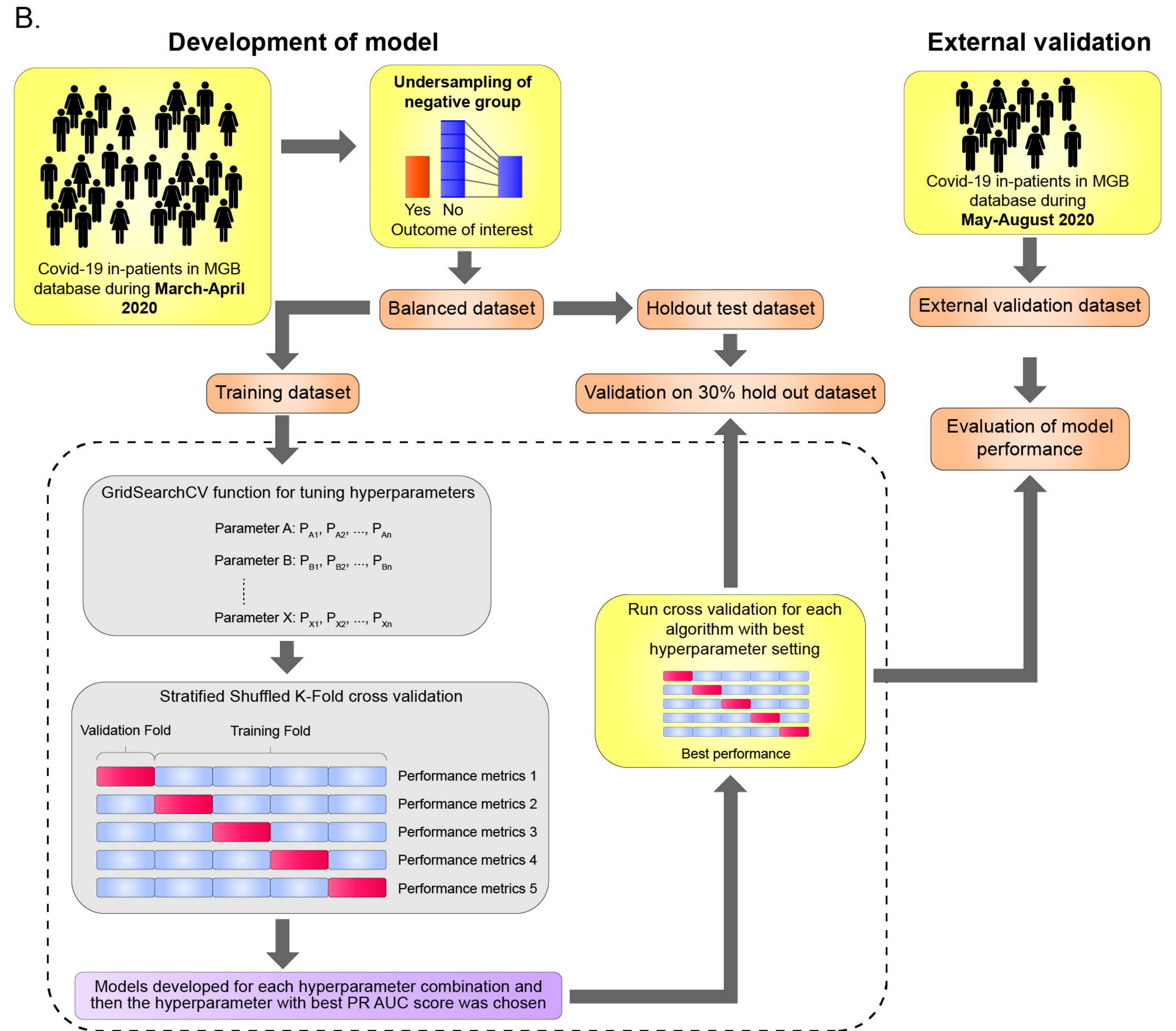
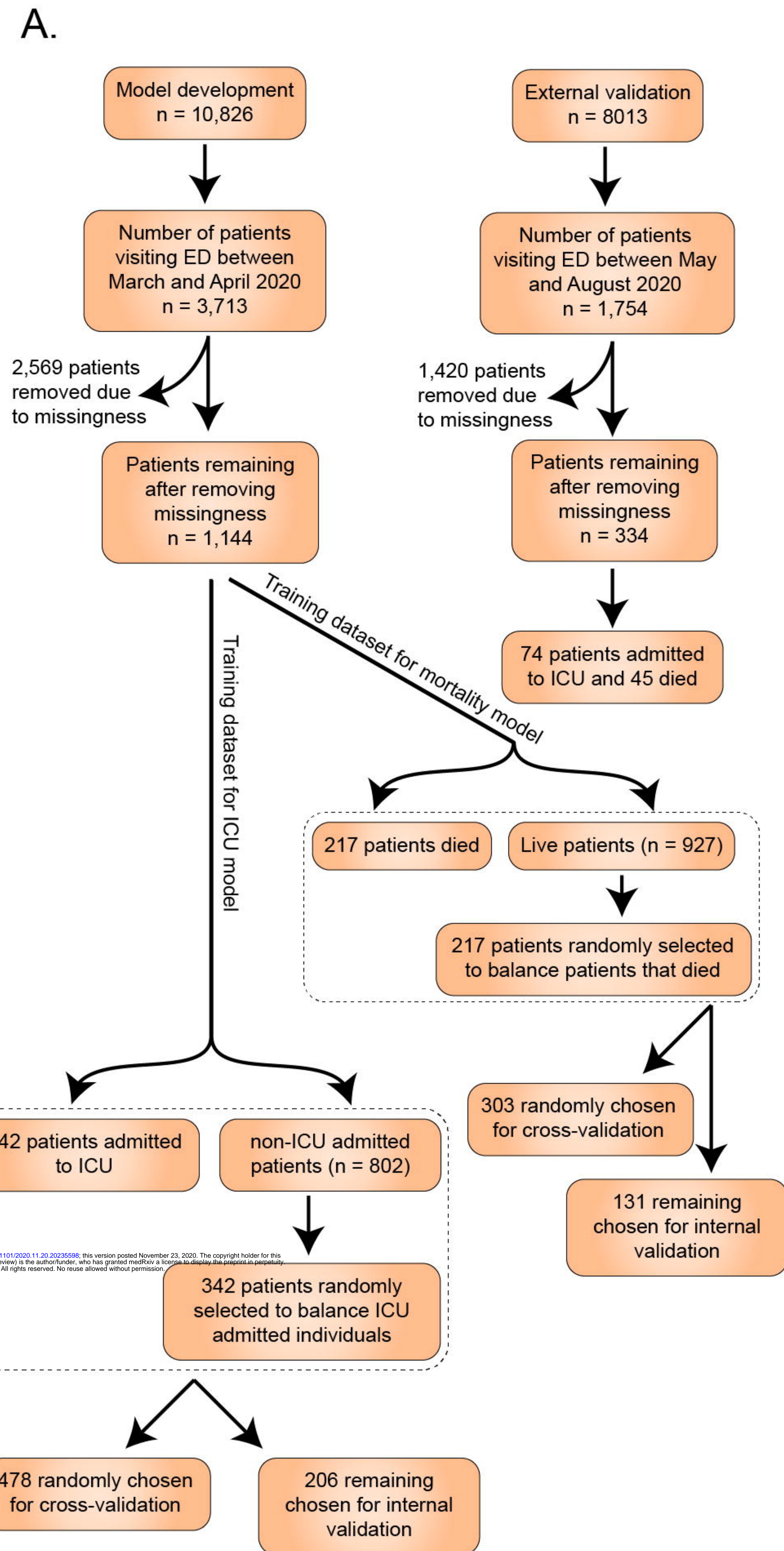
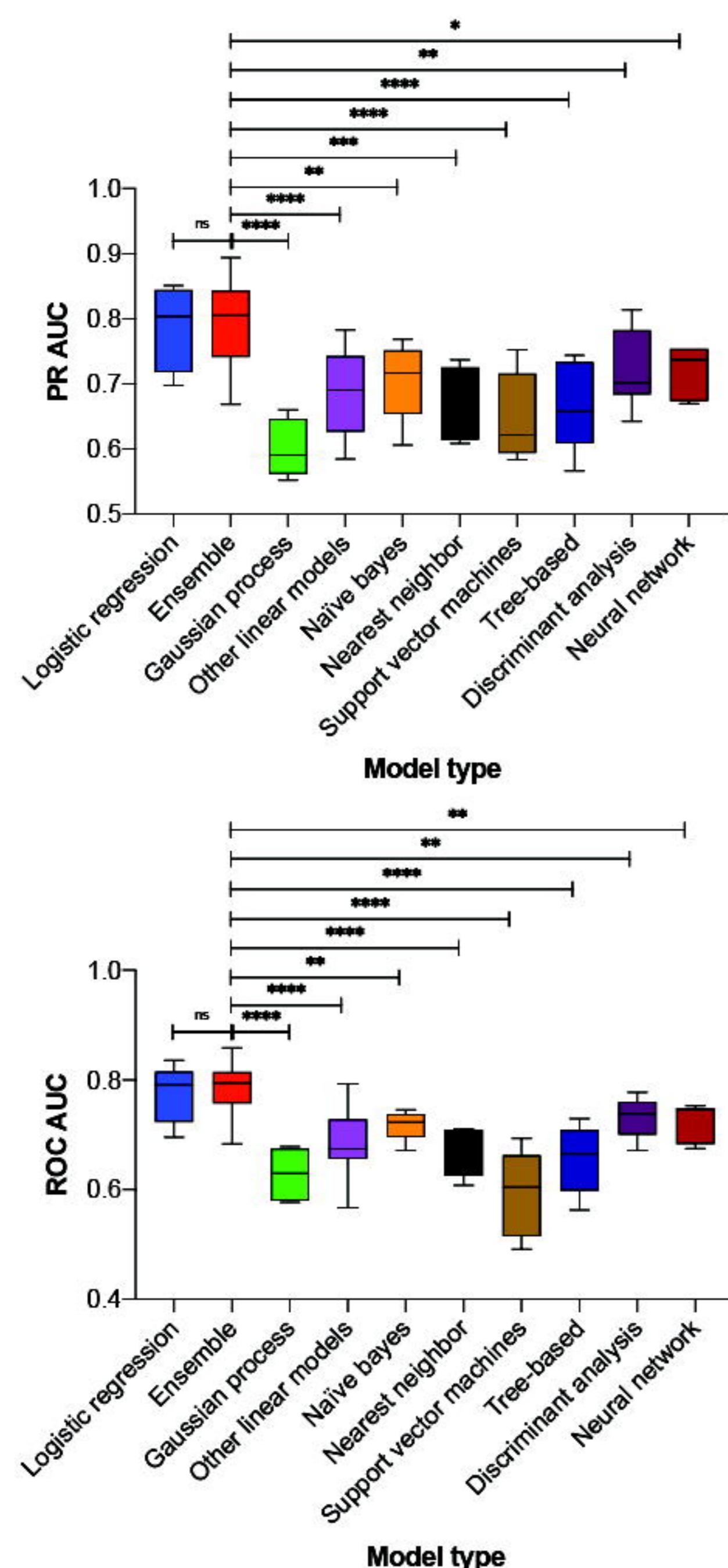
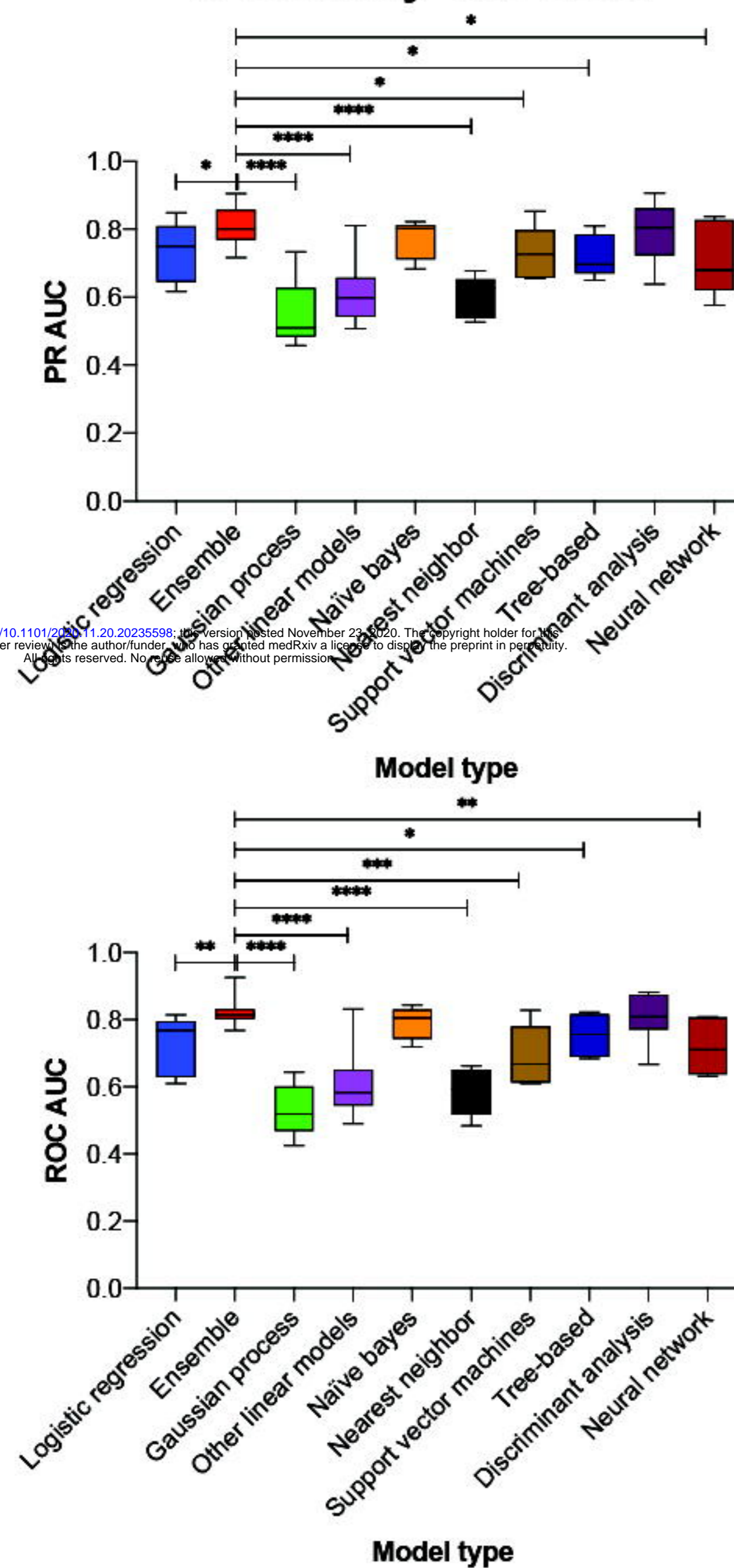


Figure 2

A. ICU admission models

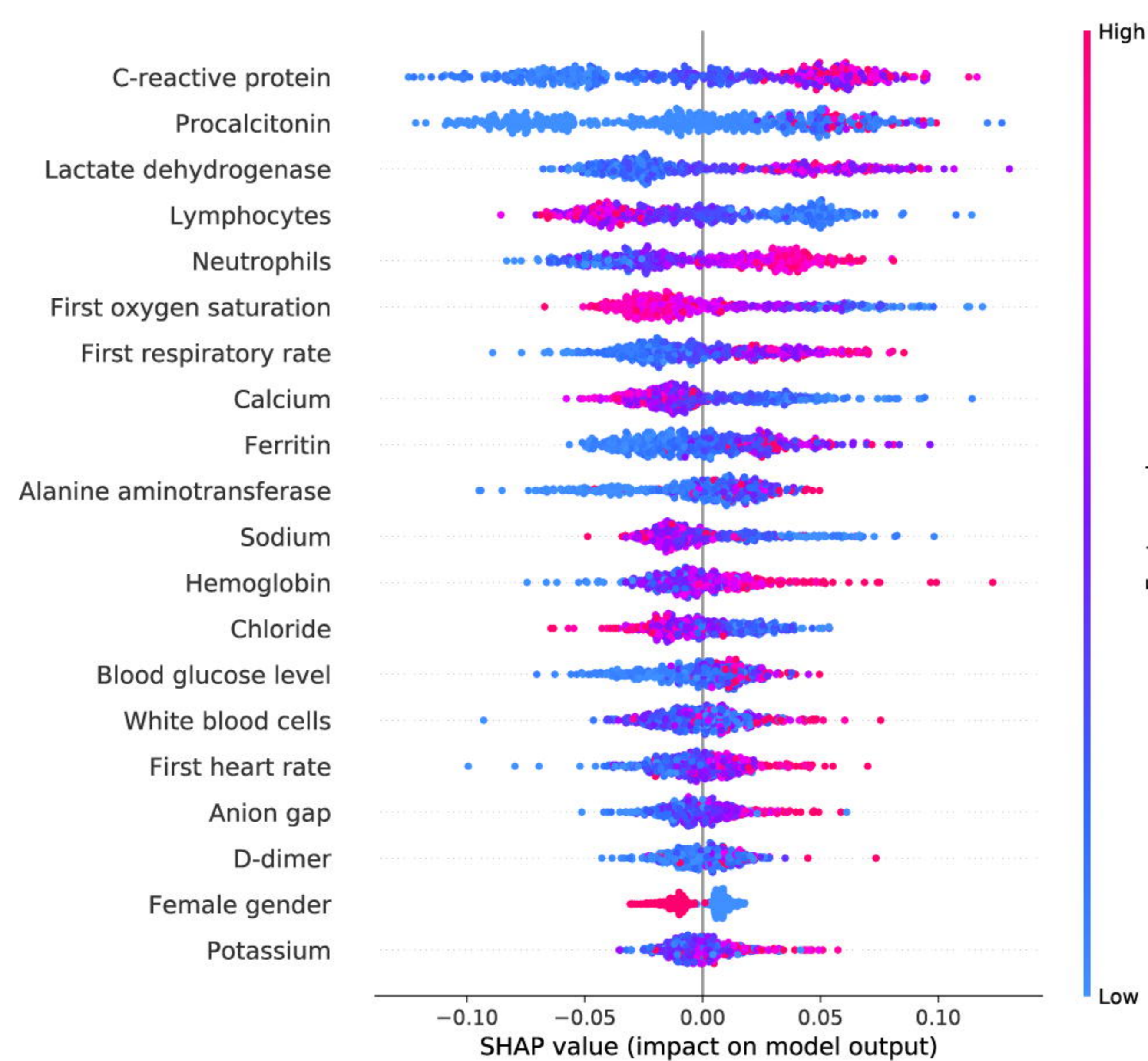


B. Mortality models

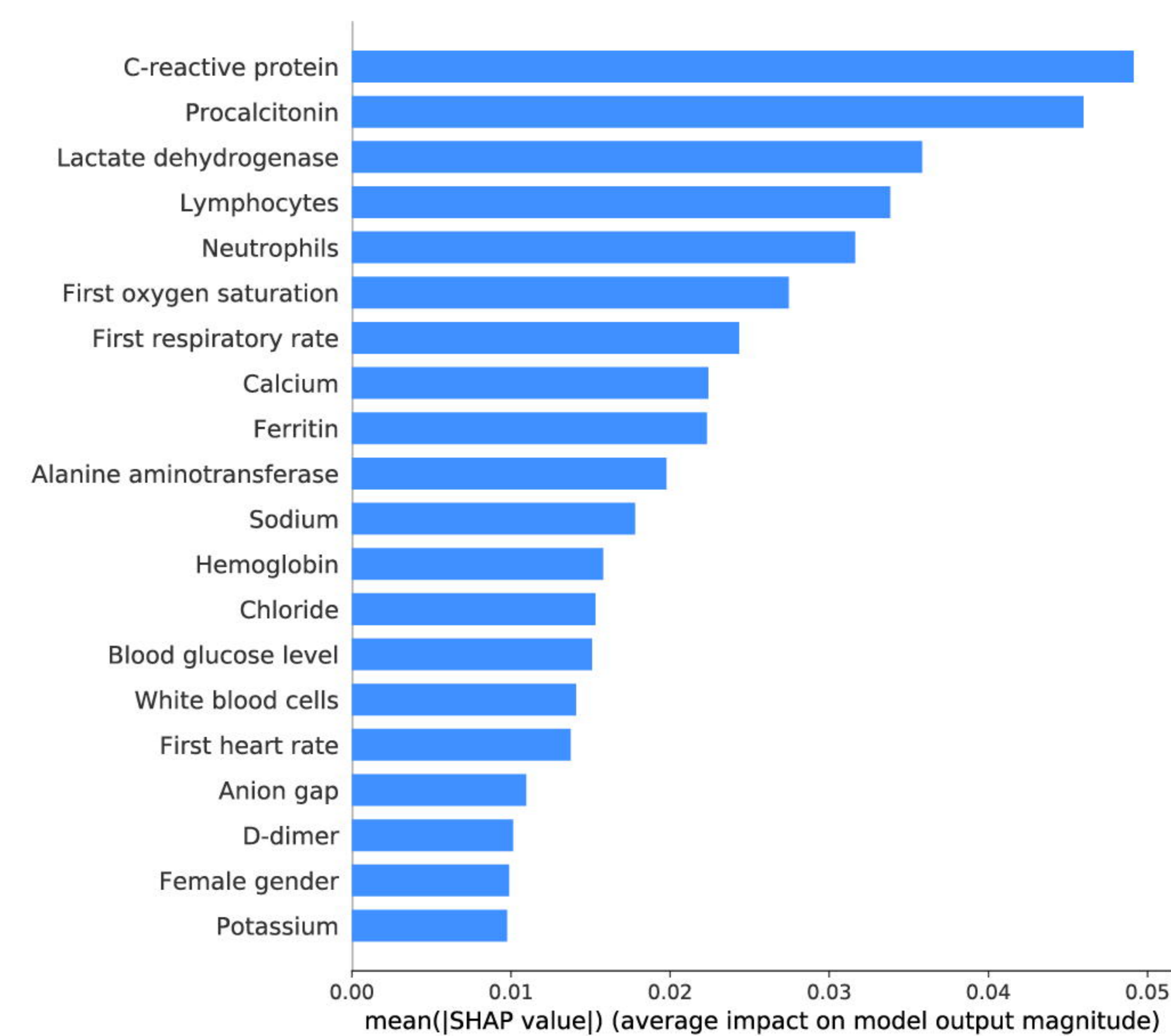


RandomForest model for predicting ICU admission

C.

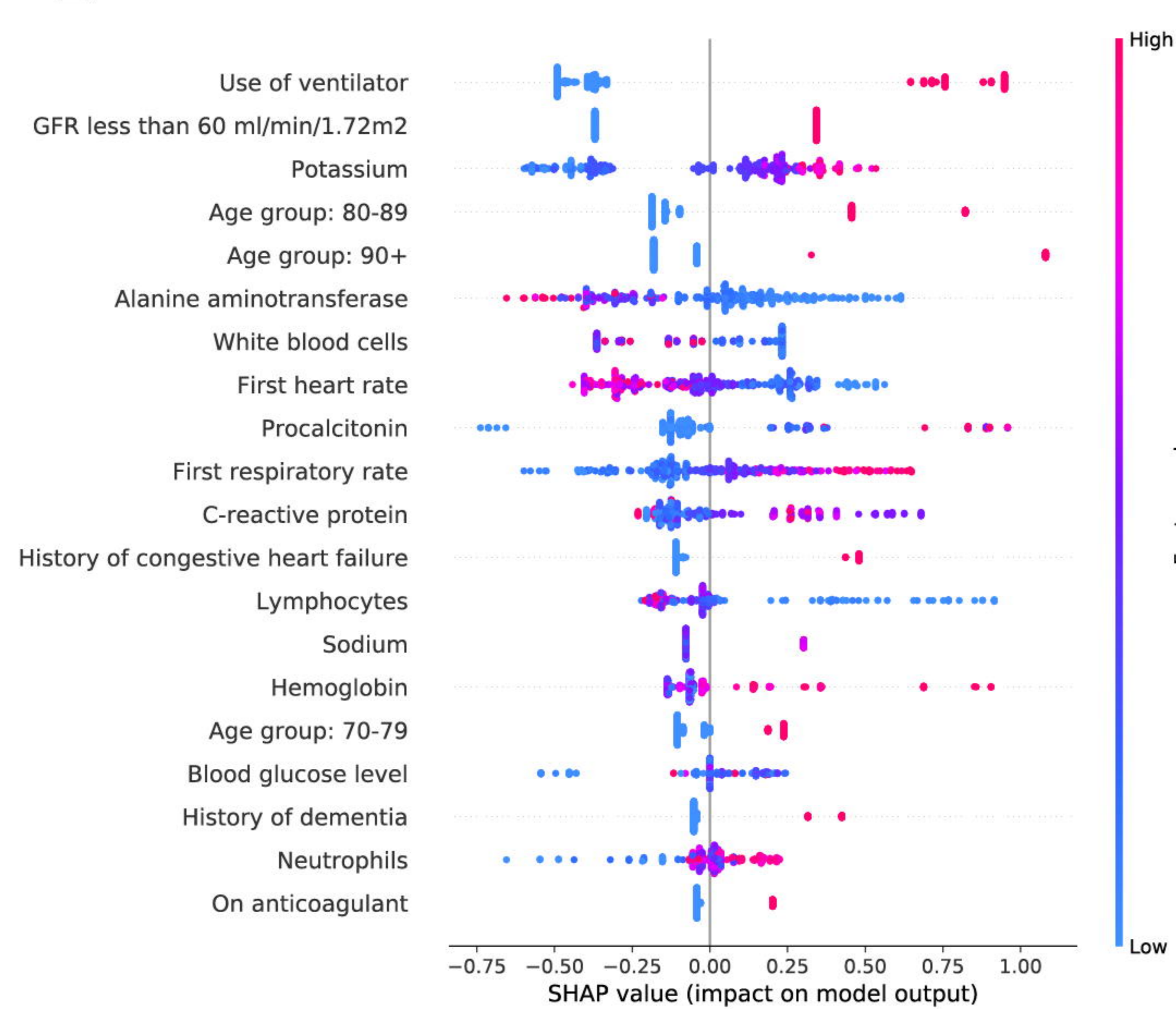


D.



XGBoost model for predicting death

E.



F.

