



Analysis of COVID-19 inpatients in France during first lockdown of 2020 using explainability methods

Jean-Baptiste Excoffier¹ · Noémie Salaün-Penquer¹ · Matthieu Ortala¹ · Mathilde Raphaël-Rousseau² · Christos Chouaid^{3,4} · Camille Jung⁵

Received: 25 October 2021 / Accepted: 23 February 2022 / Published online: 14 April 2022
© International Federation for Medical and Biological Engineering 2022

Abstract

The COVID-19 pandemic rapidly puts a heavy pressure on hospital centers, especially on intensive care units. There was an urgent need for tools to understand typology of COVID-19 patients and identify those most at risk of aggravation during their hospital stay. Data included more than 400 patients hospitalized due to COVID-19 during the first wave in France (spring of 2020) with clinical and biological features. Machine learning and explainability methods were used to construct an aggravation risk score and analyzed feature effects. The model had a robust AUC ROC Score of 81%. Most important features were age, chest CT Severity and biological variables such as CRP, O₂ Saturation and Eosinophils. Several features showed strong non-linear effects, especially for CT Severity. Interaction effects were also detected between age and gender as well as age and Eosinophils. Clustering techniques stratified inpatients in three main subgroups (low aggravation risk with no risk factor, medium risk due to their high age, and high risk mainly due to high CT Severity and abnormal biological values). This in-depth analysis determined significantly distinct typologies of inpatients, which facilitated definition of medical protocols to deliver the most appropriate cares for each profile.

Keywords COVID-19 · Machine learning · Explainable artificial intelligence · Instance selection

✉ Jean-Baptiste Excoffier
jeanbaptiste.excoffier@kaduceo.com

Noémie Salaün-Penquer
noemie.salaun-penquer@kaduceo.com

Matthieu Ortala
matthieu.ortala@kaduceo.com

Mathilde Raphaël-Rousseau
mathilderaphael-rousseau@chicreteil.fr

Christos Chouaid
christos.chouaid@chicreteil.fr

Camille Jung
camille.jung@chicreteil.fr

1 Introduction

At the beginning of the pandemic, COVID-19 was regarded as a potentially severe disease, mostly affecting elderly people requiring intensive care. Obesity and diabetes were identified as risk factors of aggravation [1, 2]. In this context, extraordinary measures were taken with respect to the care of hospitalized patients. COVID-19 patients were isolated from other patients, in separate medical units. Buildings were thoroughly cleaned and the use of protecting gear was required for all health care personnel. These measures were implemented to limit a possible contagion. Molecular microbiological determinations were used to provide an early diagnosis of the virus. Extraordinary resources were provided in order to increase intensive care unit (ICU) capacity [3, 4]. However, there was at first a clear lack of tools to identify COVID-19 patients hospitalized with a high risk of worsening condition.

Although a large amount of information about the epidemiology and clinical management of COVID-19 infection had been obtained in a remarkably short period

¹ Kaduceo, Toulouse, France
² Department of Medical Information, CHI, Créteil, France
³ Department of Pneumology, CHI, Créteil, France
⁴ Inserm U955, UPEC, IMRB, Créteil, France
⁵ Clinical Research Center, CHI, Créteil, France

[5, 6], a major gap existed in understanding the disease's severity and identifying at-risk populations at a local level. Physicians needed a clear view of hospitalized patients profiles, especially for severe cases of COVID-19. A risk score would help hospital staff identify the patients most likely to develop a severe form of COVID-19.

Several COVID-19 risk scores were thus developed [7–10] right after the outburst of the pandemic. Most studies used classical analysis approach or did not perform a thorough examination of risk score decision process when using more recent risk score construction methods such as machine learning (ML) models [11, 12]. This examination would offer the physicians a better understanding of COVID-19 medical context regarding patient stratification [13], identifying risk factors and interaction effects between relevant data points.

Aggravation risk can vary widely among patients due to important non-linearity and interaction effects. State-of-the-art ML models such as Tree Ensembles (Random Forest or Boosted Trees), Support Vector Machine or Neural Networks are now widely used in clinical context. By taking into account the differences between individuals, these models outperform the more traditional methods [14]. Boosted tree ensemble models proved to be particularly efficient in aggravation scores for COVID-19 [12, 15].

However, such ML models are more complex, making their decision process harder to understand. Several techniques have been perfected to get a better comprehension of models, grouped under the label “eXplainable Artificial Intelligence”, abbreviated as “XAI”. Such techniques produce local explanations representing, for each observations, the relative influence of every variables. Notably, local methods such as LIME (Local Interpretable Model-Agnostic Explanations) [16] or SHAP (SHapley Additive exPlanations) [17] allow to measure the influence of each feature for an individual observation representing the importance and direction of the feature effect for that particular observation. Early tools were developed during the first wave of the COVID-19 pandemic to help physicians [18, 19] obtain an aggravation risk prediction for a particular patient, along with their individual protective and risk factors.

Nevertheless, to make the most accurate prognosis, choosing the most appropriate care possible for a patient, physicians not only need the patient's specific information but also a broader view to contextualize the overall patient state. Such a context consists in visualizations about global feature importance ranking, single feature effect or interaction effect of two features.

Moreover, clustering methods can also be applied to regroup patients with similar conditions and prognosis in order to help physicians define medical procedures that are specific and most appropriate for each group of patients. A presentation of the most typical patient for each of these

homogeneous group can also facilitate the understanding of each group specificity.

2 Methodology

2.1 Study population

Population consisted in all patients hospitalized because of COVID-19 at the Centre Hospitalier Intercommunal de Créteil (CHIC) with an admission stay between the 20/03/2020 and the 09/05/2020. All hospitalizations were primarily due to COVID-19. All patients presented clear signs of COVID-19 infection leading to their hospital admission or their transfer to a unit specifically dedicated to COVID-19 if they were already hospitalized. No patient positive to COVID-19 but asymptomatic was included, since their hospitalization is only related but not directly caused by COVID-19 issues.

Data collected consisted in extractions of COVID-19 patients from the *Programme de médicalisation des systèmes d'information* (PMSI) which contains standardized medical information about hospitalizations, such as the diagnoses under the *International Classification of Diseases standard v10* (ICD-10). Data was entered on an ongoing basis by medical information technicians from the electronic medical record and checked by the physician in charge of the medical information (MR).

The primary outcome of interest was aggravation of the patient state during hospitalization. Aggravation was defined as a composite variable, thus a patient was considered aggravated if at least one of the following points was verified.

- Presence of septic shock.
- Requirements for mechanical ventilation.
- Acute respiratory distress syndrome (ARDS).
- Requirement for resuscitation maneuvers.
- Hospital mortality.

Since the period considered (spring 2020) was the initial outbreak of the COVID-19 pandemic, there was no clear guidelines about the definition of aggravation. However, even if previous studies considered slightly different definitions of aggravation [7–12], all of them included mechanical ventilation, admission to ICU and death, mostly based on previous recommendations related to pneumonia disease [20]. Moreover, physicians opted for a single binary composite target variable in order to have a clearer and faster understanding of risk factors, as done in most early studies [7, 9–11].

Every aggravation was validated by the physician in charge of the medical information (MR) using the medical record. Patient information, also called explanatory

variables, included clinical features such as age, gender, several comorbidities and results of the computerized tomography (CT) radiological exam. Additionally, several biological features, such as rates of C-reactive protein (CRP), Platelets, O2 Saturation, were also included. All explanatory variables were collected at the time of admission. Moreover, there was no missing data. For each patient, the prediction was made once, at the admission day right after the collection of the clinical and biological data forming the explanatory variables.

2.2 Statistical analysis

Statistical tests were performed in order to see the actual differences between aggravated and non-aggravated patients. Student test was used for quantitative variables (age and biological features), while the chi-square test was applied for qualitative variables (gender, comorbidities and CT Severity result). The CT Severity feature was an ordinal qualitative variable scoring lesions from best to worst, as indicated by following list :

- 0 : Scanner not needed
- 1 : Minimal
- 2 : Moderate
- 3 : Expanded
- 4 : Severe
- 5 : Critical

All *p*-values were adjusted using Bonferroni correction so as to control the family-wise error rate.

2.3 Machine learning modeling

A Boosted Tree Ensemble was used for the construction of the aggravation risk score. The XGBoost implementation was used since it is considered one of the most efficient model for tabular data [14, 15, 21].

Performances were assessed using 5-fold cross-validation. Optimization of hyperparameters was performed with an inner 5-fold cross-validation.

2.4 Influence-based analysis

A special attention was paid to the notion of explainability in order to give insights about patients profiles in respect of their personal characteristics. The SHapley Additive exPlanation (SHAP) [17, 22] was applied to observe the influence of features individually for each patient. The predictive model used being a Boosted Tree Ensemble, the TreeSHAP algorithm [22] was preferred since it is a more efficient version for tree-based models compared to its model-agnostic counterpart (both in terms of computation time and accuracy) [17].

In order to have a first view of the contribution level of each feature, a global feature importance ranking was done using the influences computed by the SHAP method. A distribution graph was then used to have a first view of the link between the feature initial value (for example an age of 80) with the related influence (that this age of 80 years old increased the aggravation probability by 30%).

Moreover, univariate graphs were constructed to offer a more detailed view about effect of a single feature, especially the most important ones selected through the global feature importance ranking. Bivariate graphs were also constructed to display the interaction effect between two features.

Clustering was also performed on the global dataset so as to group patients with similar condition and aggravation risk. The K-Medoids algorithm was used on the influences and not the initial feature values. There are indeed several advantages of using influence values rather than initial ones, as shown in Table 1.

Silhouette coefficient (also called Silhouette Score) was used to select the optimal number of clusters. The Silhouette

Table 1 Advantages of using influence values over initial values for clustering

	Initial values	Influences values
Unit	Several different units (for example age in years, oxygen saturation in percentages or a comorbidity being a binary indicator).	Single unit, the same as the prediction task (a percentage in our study since the prediction is the probability of aggravation).
Threshold	Differences are linear (it is the same between an age of 55 and 60 as between 60 and 65).	Take into account non-linearity and sudden thresholds (for example a sudden increase of the aggravation risk starting at the age of 60, meaning that a difference between an age of 60 and 65 will be much larger than between 55 and 60 in terms of influences).

Table 2 Population characteristics

		Total	Non-aggravation	Aggravation	<i>p</i> -value
Quantitative features	Nb patients	409	233 (57.0)	176 (43.0)	
	Age (years)	61.58 (±23.9)	54.77 (±25.9)	70.59 (±17.2)	<0.01 **
	Platelets (G/L)	227.48 (±93.9)	238.9 (±101.6)	212.37 (±80.5)	0.115
	Eosinophils (G/L)	0.08 (±0.2)	0.11 (±0.3)	0.05 (±0.2)	0.4075
	Neutrophils (G/L)	11.05 (±17.3)	9.8 (±14.9)	12.69 (±20.0)	1.0
	CRP (mg/L)	85.26 (±84.2)	61.98 (±68.3)	116.07 (±93.2)	<0.01 **
	O2 Saturation (%)	95.02 (±4.6)	96.28 (±2.8)	93.35 (±5.8)	<0.01 **
	Serum Creatinine (μmol/l)	95.13 (±119.6)	82.3 (±62.9)	112.12 (±166.1)	0.31
	White Globules (G/L)	20.44 (±227.1)	9.05 (±5.4)	35.52 (±346.1)	1.0
	Blood Sugar (mmol/L)	7.23 (±3.0)	6.72 (±2.1)	7.91 (±3.7)	<0.01 **
	Systolic Blood Pressure (mmHg)	132.71 (±24.1)	131.8 (±24.2)	133.92 (±24.1)	1.0
	Hemoglobin (mmol/L)	12.72 (±2.1)	12.6 (±2.1)	12.87 (±2.0)	1.0
	Arterial pulse (bpm)	96.31 (±23.6)	98.35 (±27.1)	93.6 (±17.6)	0.8175
Qualitative features	Gender	190 (46.5)	126 (54.1)	64 (36.4)	<0.05 *
	Anosmia Ageusia	362 (88.5)	196 (84.1)	166 (94.3)	0.0575
	Cancer	28 (6.8)	15 (6.4)	13 (7.4)	1.0
	Cardiovascular	141 (34.5)	75 (32.2)	66 (37.5)	1.0
	Overweight/Obesity	76 (18.6)	39 (16.7)	37 (21.0)	1.0
	Insulin Intake	40 (9.8)	15 (6.4)	25 (14.2)	0.3575
	Type 2 Diabetes	77 (18.8)	33 (14.2)	44 (25.0)	0.2025
	CT Severity: 0	180 (44.0)	123 (52.8)	57 (32.4)	<0.01 **
	CT Severity: 1	34 (8.3)	21 (9.0)	13 (7.4)	1.0
	CT Severity: 2	69 (16.9)	47 (20.2)	22 (12.5)	1.0
	CT Severity: 3	66 (16.1)	34 (14.6)	32 (18.2)	1.0
	CT Severity: 4	53 (13.0)	8 (3.4)	45 (25.6)	<0.01 **
CT Severity: 5	7 (1.7)	0 (0.0)	7 (4.0)	0.18	

Results are presented with mean and standard deviation for quantitative features, and numbers and proportion for qualitative features. All qualitative features are binary indicators except for *CT Severity*. Thus, *CT Severity* was split into six binary indicators, one for every modality *Gender* value is 0 for man and 1 for woman

For visual help for *p*-value, a single star (*) denotes a *p*-value strictly inferior to 0.05 and two stars (**) denotes a *p*-value strictly inferior to 0.01

coefficient ranges from -1 to 1 , with -1 being the worst clustering result possible and 1 the best [23, 24]. Clusters were characterized by statistical testing and presentation of their associated medoid corresponding to the most representative patient of the group, which is part of a larger problematic called *instance selection*.

2.5 Ethics

All data were anonymized before analysis, in strict observance of legislation on observational studies. Indeed, this study is compliant with the GDPR (*General Data Protection Regulation*) rules and the CNIL (*Commission Nationale de l'Informatique et des Libertés*) methodology.

The use of PMSI data for research purposes does not require an individual consent. However, patients are informed on this kind of research on the hospital's website.¹

3 Results

3.1 Population study

There was a total of 409 inpatients during the study period. Table 2 displays the principal characteristics for the global, non-aggravated and aggravated populations, such as number of patients. Statistical tests were performed between

¹<https://www.chicreteil.fr/politique-de-confidentialite/>

Table 3 Confusion matrix

Ground truth \ Prediction	Non-aggravation	Aggravation
	Non-aggravation	186
Aggravation	55	121

non-aggravated and aggravated populations. Age and several biological features came out as significantly different between aggravation and non-aggravation. As for qualitative features, Gender and several levels of CT Severity were significantly different between the two groups.

3.2 Machine learning modeling

3.2.1 Performances

Internal validation gave a global accuracy of 75.1%, with a F1-Score of 79% for the non-aggravation class and 70% for the aggravation class. Table 3 shows the confusion matrix produced by the internal validation procedure, and Table 4 indicates additional measures per class, such as sensitivity, precision and F1-Score.

Receiver operating characteristic (ROC) score was 80.7% and Precision-Recall (PR) score was 74.4%, as show by Fig. 1 that displays ROC and PR curves.

3.2.2 Influence-based analysis

Most impactful features are given by Fig. 2. Age was clearly the most important variable, followed by CT Severity. Then several biological features (CRP, O2 Saturation, Eosinophils, Platelets and Serum Creatinine) were also considered as important information by the model.

Distribution graph, also shown by Fig. 2, indicates that aggravation risk increased with Age or CT Severity. As for biological features, a higher CRP increased aggravation probability, while it was a lower O2 Saturation that increased aggravation risk.

Univariate effects of the four most important features as indicated by the feature importance ranking (Fig. 2) are displayed in Fig. 3. Age effect was mostly linear, except for young ages. CT Severity influence was similar for the first

Table 4 Performance measures per ground-truth class

Ground truth	Sensitivity	Precision	F1-score
Non-aggravation	79.8	77.2	78.5
Aggravation	68.8	72.0	70.3

three levels, then suddenly increased starting from the fourth level.

Bivariate graphs show interaction effects between two features. Figure 4 indicates interactions between Age, which was the most important feature, and the two most important qualitative features that were CT Severity and Gender. There was no clear sign of strong interaction between Age and CT Severity. But there was a slight interaction effect between Age and Gender since the difference between man and woman increased with Age.

Figure 5 shows interactions between Age and the fourth most important quantitative features after Age that were CRP, O2 Saturation, Eosinophils and Platelets. There were no clear signs of interactions between Age and CRP, O2 Saturation, or Platelets. On the contrary, there was an interaction effect between Age and Eosinophils as an older Age increased the gap between low Eosinophils values and high ones. Therefore, an older Age increased the Eosinophils impact.

3.2.3 Clustering analysis

The optimal number of clusters in respect of the Silhouette coefficient criterion was 3. Table 5 indicates the principal characteristics of clusters as well as statistical difference testing between clusters for all features. Clusters are rather homogeneous in terms of number of patients, even if the cluster 2 had the highest proportion of patients. On the contrary, clusters widely differed on the proportion of actual aggravation, with the cluster 1 having the least amount of aggravation cases, while the cluster 3 had the most.

Clusters significantly differed on patient typology. Cluster 1 had the youngest patients and cluster 2 the oldest ones. Patients of cluster 3 had more comorbidities, especially for Overweight/Obesity and Type 2 Diabetes, and more severe scanner results since the proportion of CT Severity value of 0 (scanner not done) was the lowest and that values of 4 or 5 were the highest for cluster 3. Moreover, biological feature results were also significantly worst for cluster 3, since patients of cluster 3 had the lowest O2 Saturation and Eosinophils along with the highest CRP and Serum Creatinine values.

Figure 6 shows the three medoids, which were actual observations from the dataset, corresponding to the respective three clusters. Each medoid is thus the patient most representative of a particular cluster.

Medoid of cluster 1 was the youngest, while cluster 2 medoid was the oldest. The influence of Age was the highest for cluster 1, since it decreased the aggravation probability prediction by almost 15%. Moreover, medoids of clusters 1 and 2 had very good CT Severity results (value of 0), which was a clear protective factor. On the contrary, medoid of cluster 3 had a worse CT Severity value of 3, which

Fig. 1 ROC and PR curves

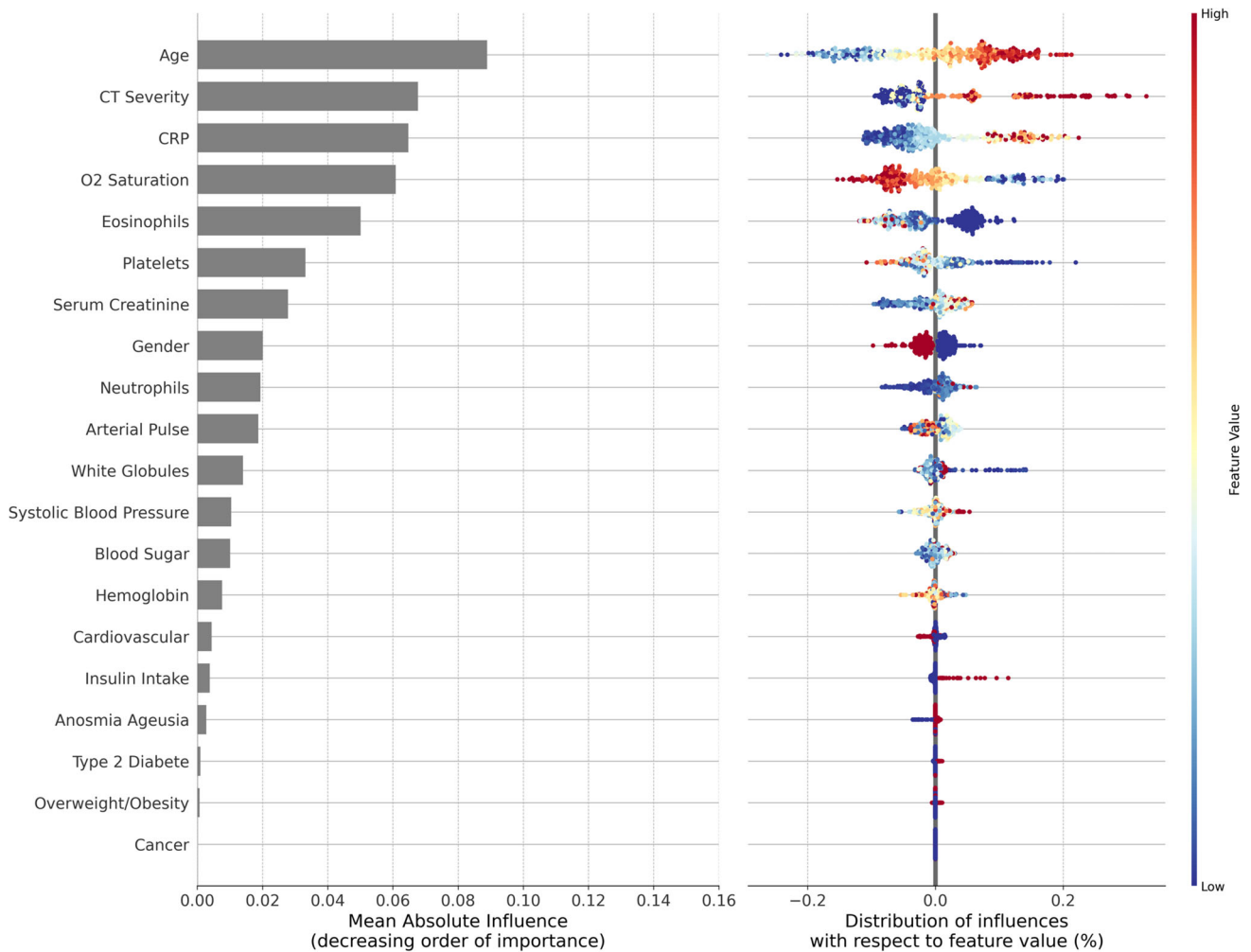
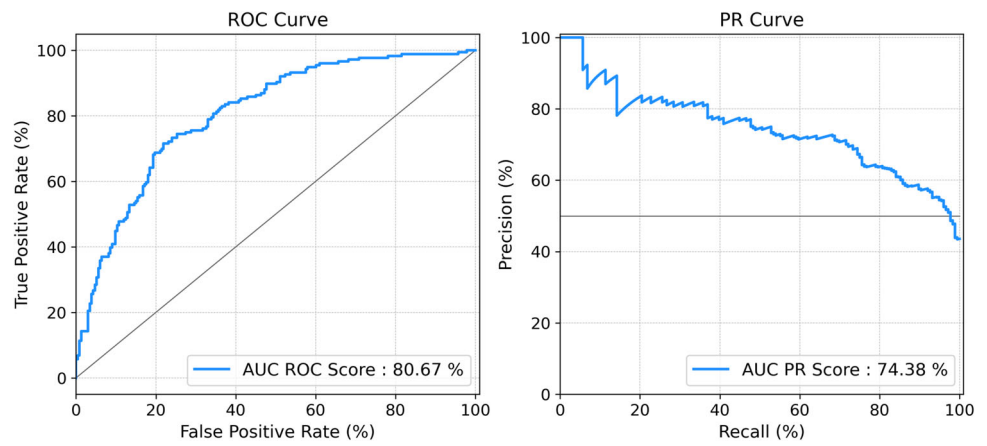


Fig. 2 Feature importance ranking and distribution of influences linked to feature initial values (best viewed in color). Feature importance ranking (left graph) is established by averaging the absolute influences over all patients. As for distribution of influences (right graph), each dot represents a particular patient, with feature names on the y-axis and influence values on the x-axis. A protective factor is

indicated by a negative influence since it decreased the probability of aggravation, while a risk factor is indicated by a positive influence. The initial value is represented by the color of the dot, through the colormap shown on the far right of the graph. For comorbidity indicators, 0 indicates the absence and 1 the presence, while genre value is 0 for man and 1 for woman

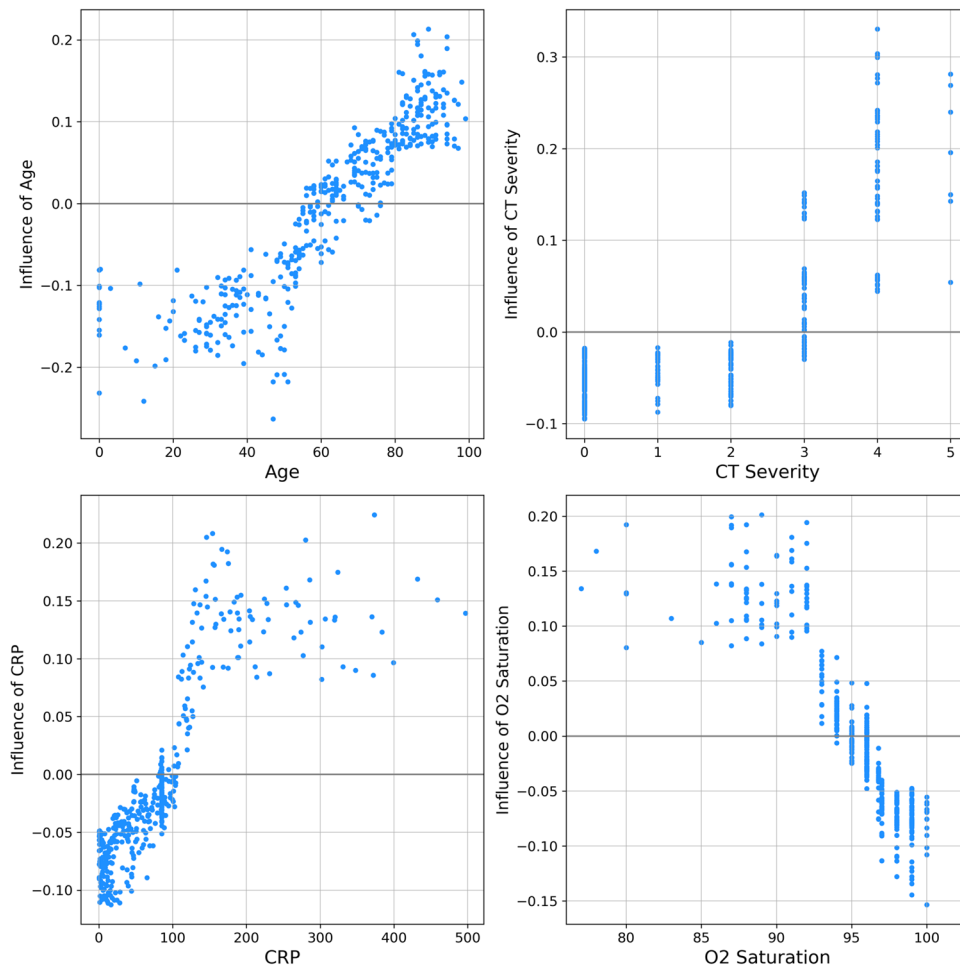


Fig. 3 Univariate graphs for effect of Age, CT Severity, CRP and O2 Saturation. Each dot represents a patient, with the feature value on the x-axis and the associated influence on the y-axis

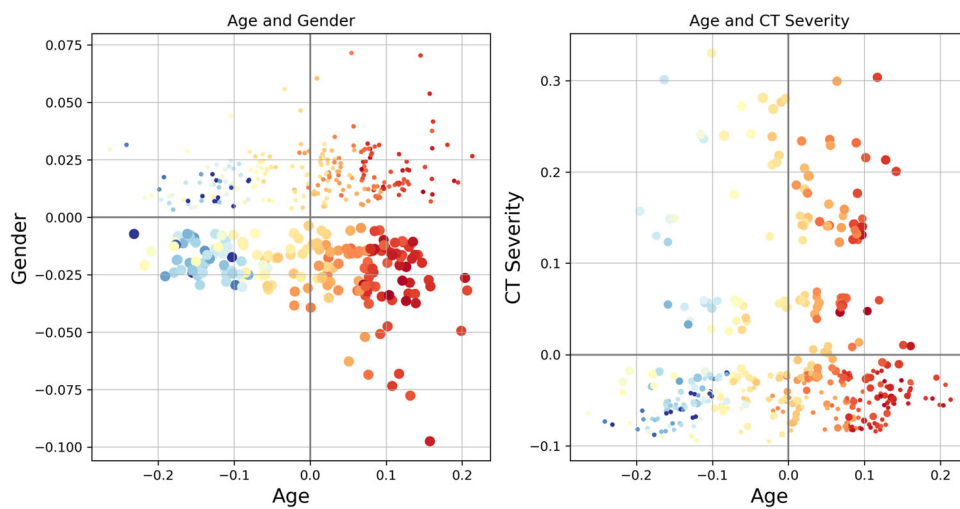
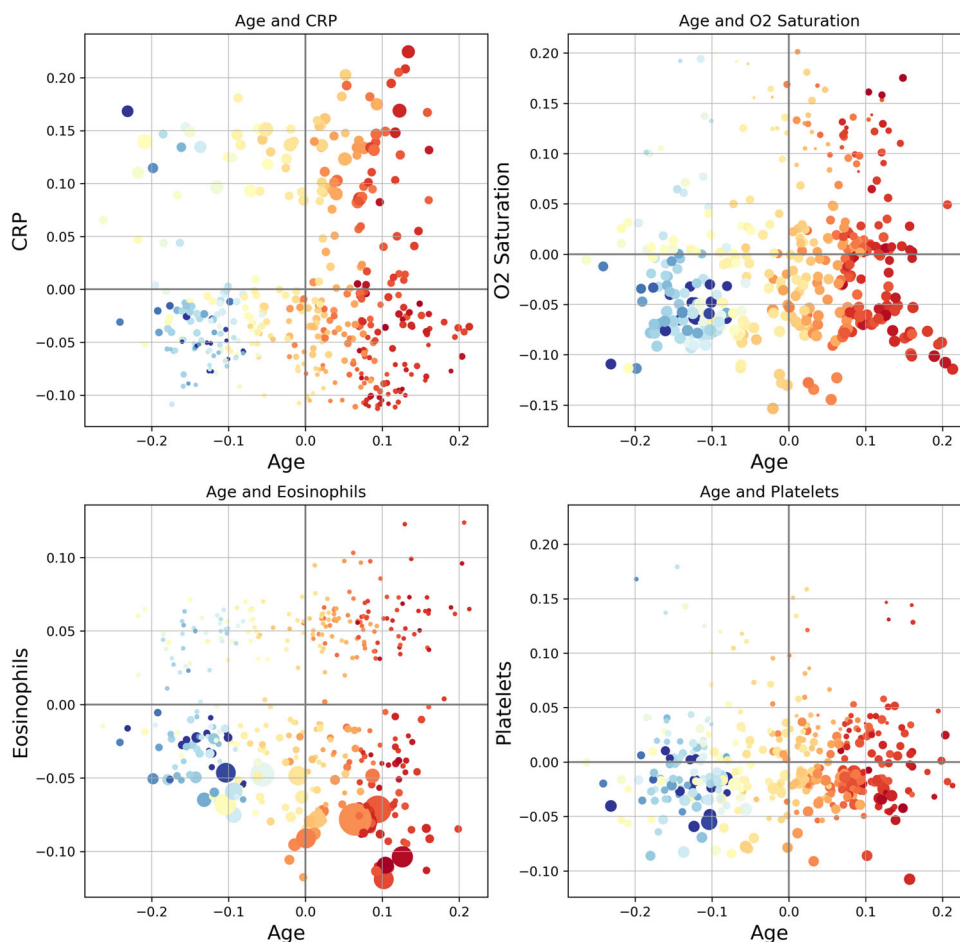


Fig. 4 Bivariate graphs for interaction effect of Age with Gender and CT Severity (best viewed in color). Each dot represents a particular patient. Color and size of the dot give information about the two feature initial values, respectively color for the feature located on the x-axis

(Age) and size for the feature on the y-axis. The colormap for the x-axis feature is the same used in Fig. 2 with blue indicating a young age and red an old age

Fig. 5 Bivariate graphs for interaction effect of Age with CRP, O2 Saturation, Eosinophils and Platelets (best viewed in color). Reading information is the same as Fig. 4



increased the predicted aggravation risk by almost 15%. Compared to the first two clusters, medoid of cluster 3 had also several biological features that were considered as risk factors, especially CRP.

4 Discussion

4.1 Main findings

This study presented the construction and evaluation of an machine learning–based aggravation risk score for COVID-19 inpatients. Though this score is based on a population from a single center, which may be considered a limitation, it did include more than 400 patients with several clinical and biological information that are available upon the patient’s arrival to the hospital.

This model showed a strong performance with an AUC ROC score of 81%. This model and the associated analysis were used as an assisting tool during the stressful context of the first wave of COVID-19 in a French hospital [18] so as to help medical staff have a better understanding of risk factors and identify population with high risk of aggravation.

The tool also presented individual protective and risk factors for each patient. However, physicians wanted and needed a broader view to get a deeper understanding of the global context. The ability to associate a single patient to a larger group and to know whether the patient is a common case, either for aggravation or non-aggravation, or a more atypical case is crucial to choose and deliver the relevant and proper care.

An in-depth analysis of the decision process of this score, thanks to machine learning and explainability methods, indicated that age, results of CT Severity scanner and several biological variables (mostly CRP, Saturation and Eosinophils [25, 26]) had a significant impact on the risk level of aggravation, which is mostly aligned with existing literature [5, 6, 27, 28].

Furthermore, more complex feature effects were highlighted by this study. Influence of CT Severity scanner result was notably non-linear, as for CRP and O2 Saturation but in a less significant manner. Interaction effect were also detected for Age and Gender, and for Age and Eosinophils [26, 29].

A clustering analysis was also performed to detect similar groups of patients based on their condition and aggravation

Table 5 Cluster characteristics

		Cluster 1	Cluster 2	Cluster 3	<i>p</i> -value
Cluster	Nb patients	129 (31.5)	172 (42.1)	108 (26.4)	
	Nb Aggravations	23 (17.8)	73 (42.4)	80 (74.1)	
	Model prediction of aggravation (%)	20.64 (14.3)	42.42 (17.7)	75.22 (13.5)	
Quantitative features	Age (years)	34.2 (±17.0)	78.49 (±11.7)	67.34 (±14.5)	<0.01 **
	Platelets (G/L)	243.42 (±106.9)	220.13 (±85.3)	220.15 (±89.0)	1.0
	Eosinophils (G/L)	0.08 (±0.2)	0.12 (±0.3)	0.03 (±0.1)	<0.01 **
	Neutrophils (G/L)	8.48 (±12.2)	10.91 (±16.8)	14.31 (±22.3)	<0.01 **
	CRP (mg/L)	59.2 (±64.5)	54.99 (±51.0)	164.59 (±96.3)	<0.01 **
	O2 Saturation (%)	96.72 (±2.5)	95.25 (±5.1)	92.63 (±4.6)	<0.01 **
	Serum Creatinine (μmol/l)	75.93 (±66.1)	96.92 (±78.1)	115.23 (±196.7)	<0.01 **
	White Globules (G/L)	9.28 (±5.6)	35.67 (±350.1)	9.51 (±4.7)	1.0
	Blood Sugar (mmol/L)	6.32 (±1.9)	6.97 (±2.3)	8.73 (±4.2)	<0.01 **
	Systolic Blood Pressure (mmHg)	126.25 (±19.0)	137.22 (±26.2)	133.23 (±24.6)	<0.05 *
	Hemoglobin (mmol/L)	12.71 (±2.3)	12.59 (±1.9)	12.93 (±2.0)	1.0
	Arterial pulse (bpm)	108.02 (±27.9)	88.19 (±19.8)	95.25 (±17.0)	<0.01 **
	Qualitative features	Gender	69 (53.5)	81 (47.1)	40 (37.0)
Anosmia Ageusia		101 (78.3)	161 (93.6)	100 (92.6)	<0.01 **
Cancer		4 (3.1)	18 (10.5)	6 (5.6)	0.8975
Cardiovascular		17 (13.2)	78 (45.3)	46 (42.6)	<0.01 **
Overweight/Obesity		27 (20.9)	24 (14.0)	25 (23.1)	1.0
Insulin Intake		7 (5.4)	19 (11.0)	14 (13.0)	1.0
Type 2 Diabetes		7 (5.4)	40 (23.3)	30 (27.8)	<0.01 **
CT Severity: 0		73 (56.6)	97 (56.4)	10 (9.3)	<0.01 **
CT Severity: 1		10 (7.8)	22 (12.8)	2 (1.9)	0.1325
CT Severity: 2		25 (19.4)	39 (22.7)	5 (4.6)	<0.01 **
CT Severity: 3		19 (14.7)	13 (7.6)	34 (31.5)	<0.01 **
CT Severity: 4	2 (1.6)	1 (0.6)	50 (46.3)	<0.01 **	
CT Severity: 5	0 (0.0)	0 (0.0)	7 (6.5)	<0.01 **	

Results are presented with numbers and proportions for *Nb patients*, *Nb Aggravations* and qualitative features, and mean and standard deviation for *Model prediction of aggravation (%)* and quantitative features

Gender value is 0 for man and 1 for woman

For visual help for *p*-value, a single star (*) denotes a *p*-value strictly inferior to 0.05 and two stars (**) denotes a *p*-value strictly inferior to 0.01

risk. Clustering was done on influence values rather than initial feature values so as to better take into account non-linearity of effects and remove the noise brought by less important features. The patients the most representative of each cluster (also called a medoid instance) were also presented so as to have a view of the different typical cases.

Three groups were identified. The first group contained patients at really low risk of aggravation represented by young patients (≤40 years), without comorbidity and with normal biological values and low CT Severity results. The second group contained a population at an intermediate aggravation risk (almost 40% of probability of aggravation) consisting of old patients (around 75 years old), but without comorbidity, and with normal biological values and low CT

Severity results. The third group had the highest aggravation risk (almost 75%) since it was comprised of patients with abnormal biological values and high CT Severity results, while age was old (around 67 years old) but not as old as the second group.

This indicated that although the age was the most important variable when considering feature independently, biological features and radiological results were clearly more impactful when all information was considered.

Therefore, this study gave a better understanding of the global situation along with information to define medical protocols based on patient characteristics so as to have pertinent guidelines that would help physicians in their work.

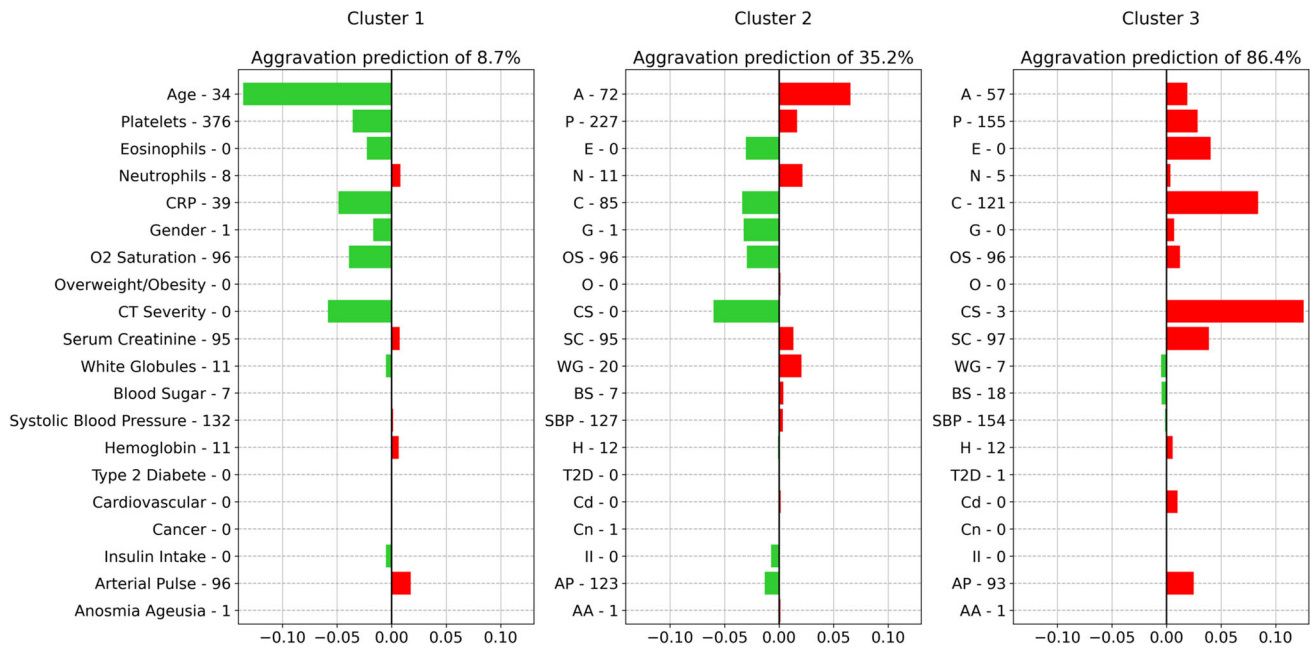


Fig. 6 Influences of patients corresponding to the medoids of the three identified clusters. Feature names are represented by their initials for cluster 2 and 3. Initial feature values are indicated after the hyphen and rounded so they all appear as integers

4.2 Future works

Even if an individual prediction for a single patient already brings an useful insight, additional information about the general context can help situate the patient in a broader view and in respect of other patients. Therefore, one could see if the patient is a typical or atypical case and bring the patient together to a larger group with similar characteristics, for which medical procedures that best fit the needs of this group have already been defined.

Identification of the most typical patient for each group, with the related characteristics along with protective and risk factors, offers a clear view of the specificity of the group. Besides, it helps physicians rapidly associate a new patient to a larger group of patients with similar condition.

This instance selection and presentation problematic was tackled in this study with a well-known, but rather classical, K-medoid technique. Testing of more recent instance selection methods, such as [30, 31], is an interesting future axis of work so as to present even more representative and useful observations to help medical staff make the most appropriate decisions.

This study focused on the outburst of the COVID-19 pandemic since it was the period with the utmost pressure on hospital centers. Still, as the model used in this study can be easily updated while admitting new patients, it could be applied to the analysis of the pandemic evolution by taking into account patients from the next waves. Indeed, very few studies have yet analyzed how the hospitalized patient profiles, especially severe cases, have evolved since

this first wave [32, 33]. This would provide an up-to-date aggravation risk score that efficiently takes into account the most recent SARS-COV-2 variants [34, 35], to help medical staff adapt their care deliveries to better reflect the current situation.

5 Conclusion

This study presented a construction of a machine learning aggravation risk score for COVID-19 inpatients. An extensive analysis based on state-of-the-art explainability techniques highlighted several non-linear and interaction effects, especially for age, CT severity result and several biological variables. Moreover, clustering method identified three main groups of patients with different risk levels. This deeper understanding of risk factors along with the obtained risk stratification can help physicians develop medical guidelines so as to improve the risk stratified care management for COVID-19 patients.

Abbreviations The following abbreviations are used in this manuscript:

CHIC, Centre Hospitalier Intercommunal Créteil; ICU, intensive care unit; ML, machine learning; XAI, eXplainable Artificial Intelligence; PMSI, Programme de médicalisation des systèmes d'information; ICD-10, International Classification of Diseases standard v10; SHAP, SHapley Additive exPlanation; GDPR, General Data Protection Regulation; CNIL, Commission Nationale de l'Informatique et des Libertés; CRP, C-reactive protein; CT, computed tomography.

Declarations

Conflict of interest The authors declare no competing interests.

References

- Stoecklin SB, Rolland P, Silue Y, Mailles A, Campese C, Simonon A et al (2020) First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures. *Eurosurveillance* 25(6):2000094
- Lampart M, Rüegg M, Jauslin AS, Simon NR, Zellweger N, Eken C et al (2021) Direct comparison of clinical characteristics, outcomes and risk prediction in patients with COVID-19 and controls—a prospective cohort study. *J Clin Med* 10(12):2021
- Institut Pasteur: projection à court terme des besoins hospitaliers pour les patients COVID-19. <https://modelisation-covid19.pasteur.fr/realtime-analysis/hospital/>
- The Latest in COVID-19 Data and Surveillance. <https://www.cdc.gov/coronavirus/2019-ncov/php/surveillance-data-analytics.html>
- Guan WJ, Liang WH, Zhao Y, Liang HR, Chen ZS, Li YM et al (2020) Comorbidity and its impact on 1590 patients with COVID-19 in China: a nationwide analysis. *Eur Respir J* 55(5)
- Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE et al (2020) Factors associated with COVID-19-related death using openSAFELY. *Nature* 584(7821):430–436
- Ji D, Zhang D, Xu J, Chen Z, Yang T, Zhao P et al (2020) Prediction for progression risk in patients with COVID-19 pneumonia: the CALL score. *Clin Infect Dis* 71(6):1393–1399
- Heldt FS, Vizcaychipi MP, Peacock S, Cinelli M, McLachlan L, Andreotti F et al (2021) Early risk assessment for COVID-19 patients from emergency department data using machine learning. *Sci Rep* 11(1):1–13
- DeCaprio D, Gartner J, Burgess T, Garcia K, Kothari S, Sayed S et al (2020) Building a COVID-19 vulnerability index. [arXiv:200307347](https://arxiv.org/abs/200307347)
- Liang W, Liang H, Ou L, Chen B, Chen A, Li C et al (2020) Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern Med* 180(8):1081–1089
- Barda N, Riesel D, Akriv A, Levy J, Finkel U, Yona G et al (2020) Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nat Commun* 11(1):1–9
- Vaid A, Somani S, Russak AJ, De Freitas JK, Chaudhry FF, Paranjpe I et al (2020) Machine learning to predict mortality and critical events in a cohort of patients with COVID-19 in New York city: model development and validation. *J Med Internet Res* 22(11):e24018
- Dera J. D. (2019) Risk stratification: a two-step process for identifying your sickest patients. *Fam Pract Manag* 26(3):21–26
- Couronné R, Probst P, Boulesteix A. L. (2018) Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinform* 19(1):1–14
- Bottino F, Tagliente E, Pasquini L, Napoli AD, Lucignani M, Figà-Talamanca L et al (2021) COVID mortality prediction with machine learning methods: a systematic review and critical appraisal. *J Pers Med* 11(9):893
- Ribeiro MT, Singh S, Guestrin C (2016) “Why should i trust you?” Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144
- Lundberg SM, Lee S. I. (2017) A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems, pp 4768–4777
- Gestions Hospitalières: Naviguer dans la tempête, n°605–April 2021. <https://kaduceo.com/naviguer-dans-la-tempete/>
- Washington University in St. Louis: towards a COVID-19 Score™ Calculator. <https://covid19score.azurewebsites.net/>
- Metlay JP, Waterer GW, Long AC, Anzueto A, Brozek J, Crothers K et al (2019) Diagnosis and treatment of adults with community-acquired pneumonia. An official clinical practice guideline of the American Thoracic Society and Infectious Diseases Society of America. *Am J Respir Crit Care Med* 200(7):e45–e67
- Chen T, Guestrin C. (2016) Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 785–794
- Lundberg SM, Erion GG, Lee SI (2018) Consistent individualized feature attribution for tree ensembles. [arXiv:180203888](https://arxiv.org/abs/180203888)
- Rousseeuw P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20:53–65
- Palacio-Niño JO, Berzal F. (2019) Evaluation metrics for unsupervised learning algorithms. [arXiv:190505667](https://arxiv.org/abs/190505667)
- Smilowitz NR, Kunichoff D, Garshick M, Shah B, Pillinger M, Hochman JS et al (2021) C-reactive protein and clinical outcomes in patients with COVID-19. *Eur Heart J* 42(23):2270–2279
- Yan B, Yang J, Xie Y, Tang X (2021) Relationship between blood eosinophil levels and COVID-19 mortality. *World Allergy Organ J* 14(3):100521
- Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW et al (2020) Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA* 323(20):2052–2059
- Grasselli G, Zangrillo A, Zanella A, Antonelli M, Cabrini L, Castelli A et al (2020) Baseline characteristics and outcomes of 1591 patients infected with SARS-cov-2 admitted to ICUs of the Lombardy Region, Italy. *JAMA* 323(16):1574–1581
- Sun C, Xue M, Yang M, Zhu L, Zhao Y, Lv X et al (2021) Early prediction of severe COVID-19 in patients by a novel immune-related predictive model. *mSphere* 6(5):e00752–21
- Kim B, Khanna R, Koyejo O. O. (2016) Examples are not enough, learn to criticize! criticism for interpretability. *Adv Neural Inf Process Syst* 29:2280–2288
- Gurumoorthy KS, Dhurandhar A, Cecchi G, Aggarwal C. (2019) Efficient data representation by selecting prototypes with importance weights. In: 2019 IEEE International conference on data mining (ICDM). IEEE, pp 260–269
- Domingo P, Poma V, Mur I, Castellvi I, Corominas H, de Benito N. (2021) Not all COVID-19 pandemic waves are alike. *Clin Microbiol Infect* 27(7):P1040.E7–1040.E10
- Jassat W, Mudara C, Ozougwu L, Tempia S, Blumberg L, Davies MA et al (2021) Difference in mortality among individuals admitted to hospital with COVID-19 during the first and second waves in South africa: a cohort study. *Lancet Global Health* 9(9):E1216–E1225
- Tregoning JS, Flight KE, Higham SL, Wang Z, Pierce B. F. (2021) Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape. *Nat Rev Immunol* 21:626–636
- Fontanet A, Autran B, Lina B, Kiény MP, Karim SSA, Sridhar D. (2021) SARS-Cov-2 variants and ending the COVID-19 pandemic. *Lancet* 397(10278):952–954

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jean-Baptiste Excoffier (MSc) is a research engineer specialized in ML applied to healthcare data with a strong focus on explicability.

Noémie Salaün-Penquer (MSc) is data science team lead at Kaduceo, expert in care pathway analysis.

Matthieu Ortala (MSc) is a computer scientist specialized in healthcare data and founder at Kaduceo.

Mathilde Raphaël-Rousseau (MD) is a physician specialized in public health, notably in medical information.

Christos Chouaid (MD PhD) is a full professor in Pneumology, board certified in Public Health.

Camille Jung (MD PhD) is a physician specialized in clinical research, notably in infectious disease.