The Influence of Inhomogeneous Input Data from Different Waves on Predictive Model Development for COVID-19 ICU Patients

Sebastian Johannes Fritsch, Konstantin Sharafutdinov, Moein Einollahzadeh Samadi, Gernot Marx, Andreas Schuppert, Johannes Bickenbach

Submitted to: Journal of Medical Internet Research
on: June 24, 2021

Table of Contents

The Influence of Inhomogeneous Input Data from Different Waves on Predictive Model Development for COVID-19 ICU Patients

Sebastian Johannes Fritsch[1, 2, 3*] MD, Dr med; Konstantin Sharafutdinov[3, 4, 5*] MSc; Moein Einollahzadeh Samadi[4, 5] MSc; Gernot Marx[1, 3] FRCA, MD, Prof Dr; Andreas Schuppert[3, 4, 5*] Prof Dr; Johannes Bickenbach[1, 3*] MD, Prof Dr

[1]Department of Intensive Care Medicine University Hospital RWTH Aachen Aachen DE
[2]Jülich Supercomputing Centre Forschungszentrum Jülich Jülich DE
[3]SMITH Consortium of the German Medical Informatics Initiative Leipzig DE
[4]Institute for Computational Biomedicine RWTH Aachen University Aachen DE
[5]Joint Research Center for Computational Biomedicine RWTH Aachen University Aachen DE
[*]these authors contributed equally

Corresponding Author:
Sebastian Johannes Fritsch MD, Dr med
Department of Intensive Care Medicine
University Hospital RWTH Aachen
Pauwelsstr. 30
Aachen
DE

Abstract

Background: During the course of the COVID-19 pandemic, a variety of machine learning models were developed to predict different aspects of the disease, such as long-term causes, organ dysfunction or ICU mortality. The number of training datasets used has increased significantly over time. However, these data now come from different waves of the pandemic, not always addressing the same therapeutic approaches over time as well as changing outcomes between two waves. The impact of these changes on model development has not yet been studied.

Objective: The aim of the investigation was to examine the predictive performance of several models trained with data from one wave predicting the second wave´s data and the impact of a pooling of these data sets. Finally, a method for comparison of different datasets for heterogeneity is introduced.

Methods: We used two datasets from wave one and two to develop several predictive models for mortality of the patients. Four classification algorithms were used: logistic regression (LR), support vector machine (SVM), random forest classifier (RF) and AdaBoost classifier (ADA). We also performed a mutual prediction on the data of that wave which was not used for training. Then, we compared the performance of models when a pooled dataset from two waves was used. The populations from the different waves were checked for heterogeneity using a convex hull analysis.

Results: 63 patients from wave one (03-06/2020) and 54 from wave two (08/2020-01/2021) were evaluated. For both waves separately, we found models reaching sufficient accuracies up to 0.79 AUROC (95%-CI 0.76-0.81) for SVM on the first wave and up 0.88 AUROC (95%-CI 0.86-0.89) for RF on the second wave. After the pooling of the data, the AUROC decreased relevantly. In the mutual prediction, models trained on second wave´s data showed, when applied on first wave´s data, a good prediction for non-survivors but an insufficient classification for survivors. The opposite situation (training: first wave, test: second wave) revealed the inverse behaviour with models correctly classifying survivors and incorrectly predicting non-survivors. The convex hull analysis for the first and second wave populations showed a more inhomogeneous distribution of underlying data when compared to randomly selected sets of patients of the same size.

Conclusions: Our work demonstrates that a larger dataset is not a universal solution to all machine learning problems in clinical settings. Rather, it shows that inhomogeneous data used to develop models can lead to serious problems. With the convex hull analysis, we offer a solution for this problem. The outcome of such an analysis can raise concerns if the pooling of different datasets would cause inhomogeneous patterns preventing a better predictive performance.

(JMIR Preprints 24/06/2021:31539)
DOI: https://doi.org/10.2196/preprints.31539

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Sebastian Johannes Fritsch, Dr, MD[a,b,c] *; Konstantin Sharafutdinov, MSc[c,d,e] *; Moein Einollahzadeh Samadi, MSc[d,e]; Gernot Marx, Prof Dr, MD[a,c]; Andreas Schuppert, Prof Dr, PhD[c,d,e] #, Johannes Bickenbach, Prof Dr, MD[a,c] #

[a]Department of Intensive Care Medicine, University Hospital RWTH Aachen, Pauwelsstr. 30, 52074 Aachen, Germany
[b]Juelich Supercomputing Centre, Forschungszentrum Juelich, Wilhelm-Johnen-Straße, 52428 Jülich, Germany
[c]SMITH Consortium of the German Medical Informatics Initiative, Leipzig, Germany
[d]Institute for Computational Biomedicine, RWTH Aachen University, Pauwelsstr. 19, 52074 Aachen, Germany
[e]Joint Research Center for Computational Biomedicine, RWTH Aachen University, Pauwelsstr. 19, 52074 Aachen, Germany

* First authors contributed equally
# Senior authors contributed equally

Corresponding author:
Sebastian Fritsch, MD
Department of Intensive Care Medicine
University hospital RWTH Aachen
Pauwelsstr. 30
52074 Aachen
sfritsch@ukaachen.de
Phone: +49-241-8080444
Fax: +49-241-8082406

The Influence of Inhomogeneous Input Data from Different Waves on Predictive Model Development for COVID-19 ICU Patients

Abstract

Background: During the course of the COVID-19 pandemic, a variety of machine learning models were developed to predict different aspects of the disease, such as long-term causes, organ dysfunction or ICU mortality. The number of training datasets used has increased significantly over time. However, these data now come from different waves of the pandemic, not always addressing the same therapeutic approaches over time as well as changing outcomes between two waves. The impact of these changes on model development has not yet been studied.
Objective: The aim of the investigation was to examine the predictive performance of several models trained with data from one wave predicting the second wave´s data and the impact of a pooling of these data sets. Finally, a method for comparison of different datasets for heterogeneity is introduced.
Methods: We used two datasets from wave one and two to develop several predictive models for mortality of the patients. Four classification algorithms were used: logistic regression (LR), support vector machine (SVM), random forest classifier (RF) and AdaBoost classifier (ADA). We also performed a mutual prediction on the data of that wave which was not used for training. Then, we

compared the performance of models when a pooled dataset from two waves was used. The populations from the different waves were checked for heterogeneity using a convex hull analysis.
Results: 63 patients from wave one (03-06/2020) and 54 from wave two (08/2020-01/2021) were evaluated. For both waves separately, we found models reaching sufficient accuracies up to 0.79 AUROC (95%-CI 0.76-0.81) for SVM on the first wave and up 0.88 AUROC (95%-CI 0.86-0.89) for RF on the second wave. After the pooling of the data, the AUROC decreased relevantly. In the mutual prediction, models trained on second wave´s data showed, when applied on first wave´s data, a good prediction for non-survivors but an insufficient classification for survivors. The opposite situation (training: first wave, test: second wave) revealed the inverse behaviour with models correctly classifying survivors and incorrectly predicting non-survivors. The convex hull analysis for the first and second wave populations showed a more inhomogeneous distribution of underlying data when compared to randomly selected sets of patients of the same size.
Conclusions: Our work demonstrates that a larger dataset is not a universal solution to all machine learning problems in clinical settings. Rather, it shows that inhomogeneous data used to develop models can lead to serious problems. With the convex hull analysis, we offer a solution for this problem. The outcome of such an analysis can raise concerns if the pooling of different datasets would cause inhomogeneous patterns preventing a better predictive performance.

Keywords: COVID-19; machine learning; model development; predictive models; mortality prediction; convex hull analysis; intensive care; mechanical ventilation

Introduction

The COVID-19 pandemic resulted in nearly overwhelmed medical systems in several countries. In this unclear situation, there was strong need for help to allocate the scarce health care resources appropriately, especially in the field of acute critical care.

In view of these needs and of previous encouraging implementations of artificial intelligence (AI) [1], many researchers developed COVID related prediction models or scores, which focused on different aspects of disease progression, like mortality or admission to an Intensive Care Unit (ICU). However, the growing number of models in that field was also met with opposition. Although almost all published models reported a high predictive performance, Wynants et al. pointed out that many of them showed some serious weaknesses, like unclear or high risks of bias, high risk of overfitting or simply an insufficient reporting [2].

Attempts to apply models developed in a single hospital onto patients from another hospital have already revealed their limitations [3]. How the performance of such models is affected by the temporal separation of the integrated population during a pandemic, e.g. when models developed during the first wave of COVID 19 are applied onto patients from following waves, even within one hospital, remains unclear. The second important question is whether it is generally legit to pool the data from the different waves of disease to develop predictive models. In newest publications, the models integrate several tens of thousands of patients pooled from the whole period of the pandemic [4]. This rather ignores the dynamical changes in therapy and subsequently in the outcome of patients. Due to the growing evidence and expertise of physicians in the therapy of the disease, the ratio of hospitalized patients who were admitted to the ICU and who required mechanical ventilation (MV) decreased [5, 6]. Latest analyses show a decreasing mortality of COVID 19 patients over the time [7-9]. These changes give the waves their respective properties.

Data-driven models, such as machine learning methods, aim to represent systems solely from available measurement data. However, a critical issue of such models is their limited extrapolability. Unless strong assumptions are posed on the learned function, data-driven models, not depending on the output to be predicted, can only be valid in regions where they have sufficiently dense coverage of training data points, which is referred to as the validity domain [10]. This can be approximated by the convex hull spanned by the data, which represents an upper bound of the validity domain for any machine learning application.
Hence, one possible approach to examine different populations for homogeneity with respect to predictive performance of machine learning models is to perform a convex hull analysis of the available data to be used for training and prediction, respectively [11, 12].

Convex hull of a set of data points is defined as the smallest polytope with dimensionality equal to the number of attributes containing the points in such way that every straight line connecting a pair of points lies inside the polytope [13, 14]. As the intersection of the convex hulls of training and test set is an upper bound for the generalization of any machine learning base model not depending on the outcome to be predicted. In the case of learning from different populations, the intersection of the convex hulls can serve as a measure for sufficient similarity of heterogeneous populations enabling the reliability of generalization of machine learning models. Hence, although machine learning models, trained on a data set A, may be predictive on a data set B even outside the intersection of the convex hulls A and B, predictivity cannot be assessed from data of A alone.

For this study, we analyzed the mutual predictivity of machine learning models developed for COVID-19 patients requiring ICU treatment including MV, who form the most severely affected group of patients. The aim of this investigation was firstly to examine how mortality prediction models developed on the data of one wave of the pandemic would perform on patients of another wave. Secondly, we explored the influence of pooling two datasets from different pandemic waves and its impact on predictive performance. Finally, we performed a convex hull analysis to examine the included populations for homogeneity and to reveal the influence of the different population structure onto the developed prediction models.

Methods

Ethics approval and Data sources

This analysis was approved by the local ethical review board (EK 091/20; Ethics Committee, Faculty of Medicine, RWTH Aachen, Aachen, Germany). The Ethics Committee waived the need to obtain Informed consent for the collection, analysis and publication of the retrospectively obtained and depersonalized data. All methods were carried out in accordance with relevant guidelines and regulations.

Data were retrieved from an electronic patient data recording system (medico//s, Siemens, Germany) and from an online patient data documentary system (IntelliSpace Critical Care and Anesthesia, ICCA Rev. F.01.01.001, Philips Electronics, Netherlands). The study cohort included severely ill patients with confirmed COVID-19, who were admitted to the ICUs at the University hospital RWTH Aachen and required MV throughout their ICU stay. For each patient, demographic, clinical and outcome data were available. All evaluated parameters were used in daily clinical routine and no additional data were collected. Data consisted from static data, like biometrics and ICD-10 codes and dynamic data including all types of clinical data, e.g. vital signs, MV parameters, laboratory parameters and medication. Invasive MV was defined by evidence of positive-end-expiratory

pressure (PEEP) and end-inspiratory pressure (P EI). Timepoint of ICU admission was defined by the first available heart rate measurement.

Data preparation and cleaning

An experienced ICU physician (S.F.) reviewed and cross-checked the data before the main analysis. The mean values of every routinely charted ICU parameter collected over the first day of ICU stay were extracted as potential predictor variables, later on also referred to as features. The total Sequential organ failure assessment (SOFA) score was calculated as described before [15] and included as a potential predictor variable, along with SOFA scores for subsystems and variables used for the calculation of these. The SOFA subscore for central nervous system was excluded, since there are no clear standards for its application in sedated MV patients [16]. Variables with values missing in more than 20 % of patients as well as features with a correlation coefficient between measurements greater than 0.95 were omitted from the analysis.

Prediction Model Development

Four different common classification algorithms were used as candidate classifiers: logistic regression (LR), support vector machine (SVM), random forest classifier (RF) and AdaBoost classifier (ADA). Death in the ICU was chosen as primary endpoint.

Due to the small sample size of the cohorts under consideration, a splitting of data into development and validation sets for the model development and tuning of hyperparameters was not feasible. Therefore, a nested cross-validation (CV) strategy was used for the model development. Under this method, choice of features in the model (complexity) and hyperparameters was performed as follows: after a nested CV procedure, an ensemble of best-performing models and their performances was available for construction of a consensus model. Features found in the majority of best models, defined the final complexity of the model. Subsequently, we defined hyperparameters of the model with a fixed complexity using the CV on the entire dataset. Lastly, the hyperparameters found during this final search were used to configure a final model, which was then fit on the entire dataset. Overall procedure was performed for four algorithms mentioned above, each providing a final model with fixed complexity and hyperparameters.

Statistical Analysis

The performance of the developed models was assessed using the area under the receiver-operator characteristic curve (AUROC), with 95 % confidence intervals calculated by repeated nested-CV procedure (20 times). Statistical analysis was performed with scikit-learn library for Python3 programming language [17, 18]. For the pairwise comparison of two sets a two-tailed Student's t-test with a significance level of α =0.05 was used.

Convex hull analysis

The intersection of the convex hulls of the first and second wave COVID-19 patients across different attribute spaces with various dimensions was evaluated. The results of the two waves were compared to those of two populations with the same sizes as the first and second wave population, but that were randomly selected from the pooled data. The two distributions of convex hull intersections were compared using a Student´s t-test with a significance level of α =0.05. Additionally, the convex hull intersection between the first and second wave populations for the fixed attribute spaces

(complexity) used in the developed consensus models was analyzed.

Results

Characteristics of the development cohorts

For model development, we used three cohorts of patients: patients of the first and the second wave of the pandemic and a pooled cohort, which comprised patients of both waves.
Patient data were collected between March and end of June 2020 for the first wave and between end of August 2020 and beginning of January 2021 for the second wave. The initial cohorts included 65 and 55 MV patients, correspondingly. After exclusion of patients with ICU stay shorter than 24 hours, who did not contribute a relevant amount of data points, the cohorts resulted in 63 patients and 54 patients, respectively.
The key characteristics of the different populations are given in Table 1. While the analyzed groups showed only slight differences in age, gender distribution and mortality rate, the duration of treatment on the ICU was nearly halved from first to second wave. Also some other clinical features of wave two cohort, like a lower PEEP, P EI, p_aCO_2 and SOFA Score, indicated a decreasing severity of disease in ICU patients of this cohort. A full list of clinical features is available in Table S1 in the Multimedia appendix.

Table 1: Biometrics and clinical data for the three analyzed patient cohorts

	Wave 1	**Wave 2**	**Pooled**
	n (%) or median (IQR)		
No.	63 (100)	54 (100)	117 (100)
Age (years)	62 (12)	60.5 (14.7)	62 (13)
Male gender	42 (67)	39 (72.2)	81 (69.2)
Mortality	27 (42.9)	30 (55.6)	57 (48.7)
Length of ICU stay (days)	27.0 (34.5)	13.6 (22.4)	20.5 (28.7)
Length of MV (days)	23.6 (30.6)	13.3 (22.3)	17.7 (26.6)

Predictive performance of the single consensus models

For each method and for every cohort, a consensus model with fixed complexity was found using the nested CV procedure. Thus, this procedure resulted in four models each for the first wave data, the second wave data and the pooled cohort. Performances of all models are listed in Table S2 in the Multimedia appendix. The number of features in the models ranged from 2 to 11 (see Table S3 and S4 in the Multimedia appendix). For wave one and wave two separately, we found models reaching intermediate to good accuracies: up to 0.79 mean AUROC for the SVM on the first wave and up 0.88 mean AUROC for the RF on the second wave.

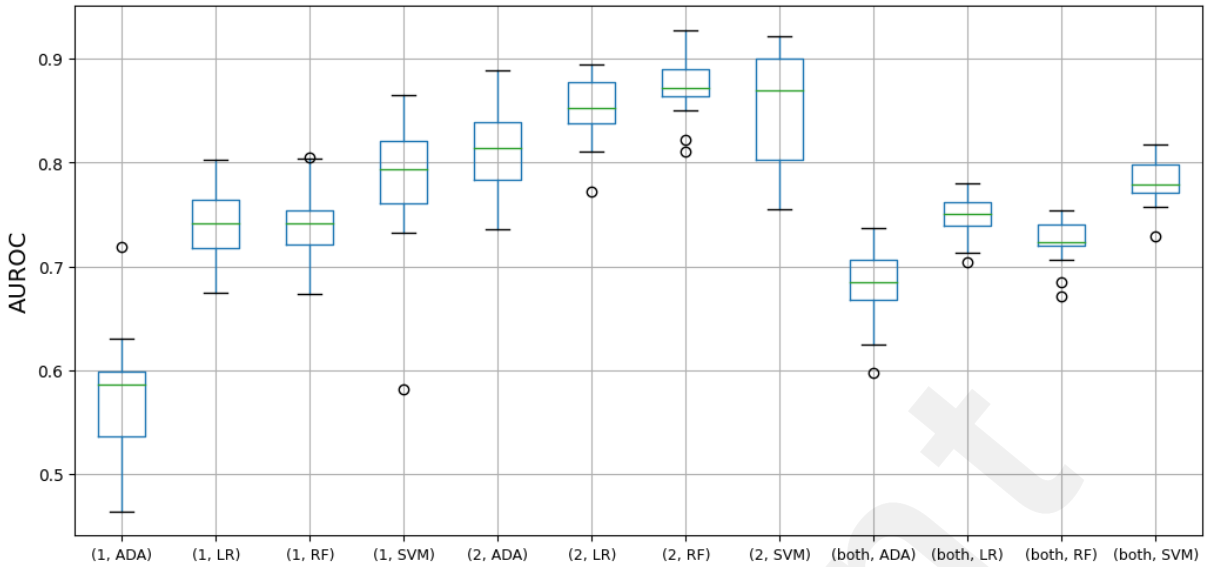


Figure 1: Performance of developed models with fixed complexity. Performance is depicted using AUROC in the nested CV procedure. From left to right: four boxplots show performances of the models developed and tested with data from the first wave only, the second wave only and the pooled data from both waves, respectively. (1: first wave data, 2: second wave data, both: pooled data, SVM: support vector machine, LR: logistic regression, RF: random forest, ADA: AdaBoost.) Boxes indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.

The SVM with radial basis function kernel achieved the best performance in terms of AUROC for first wave and the pooled cohort (see Figure 1). LR and RF performed on a similar level, but clearly inferior to the SVM. In the second wave cohort, the SVM and RF performed on a similar level, but the SVM showed a much higher Interquartile ratio (IQR), indicating a higher variation of results. Differences in predictive performance between the first and the second wave in terms of AUROC range from 0.07 (the SVM classifier) to 0.23 (ADA). Generally, ADA always performed poorer relatively to other models.

Performance of models on pooled data (Figure 1, marked with ”both”) significantly dropped (See Table S5 in Multimedia appendix) when being compared to single cohorts (Figure 1, marked with ”1” and “2”) with exceptions for ADA, which failed in the first cohort. The overall drop in AUROC ranged from 0.07 (SVM) to 0.15 (RF) when comparing models developed on pooled data to those developed on the wave two. For the wave one the insignificant changes were observed for the LR (+0.01) and the SVM (-0.003). The increase was observed for the ADA (+0.11) and decrease for the RF (-0.02).

After fixing the complexity of the developed model (features), adding data from the other cohort and checking new nested-CV results, we observed a significant decrease in the AUROC (see Tables S6.1 and S6.2 in the Multimedia appendix for detailed information on performance of the models and Table 7 for the comparison of performance). The comparisons between the performances of the models, which were developed and tested using the first and second wave cohort only and those developed and tested on pooled data are shown in the Figure 2. Models developed on both waves show similar behaviour. The drop in the performance for the first wave ranges from 0.02 for the LR to 0.1 for the SVM, with an only exception for first wave ADA. For the second wave performance decrease spans from 0.09 for the ADA to 0.15 for the RF.

Fig. 2a: Wave 1 vs. pooled data

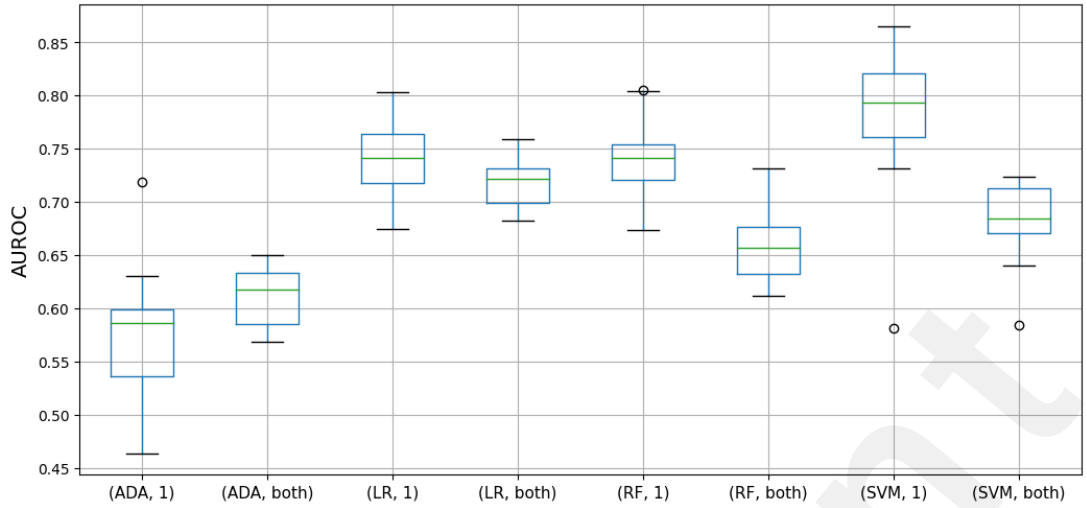


Fig. 2b: Wave 2 vs. pooled data

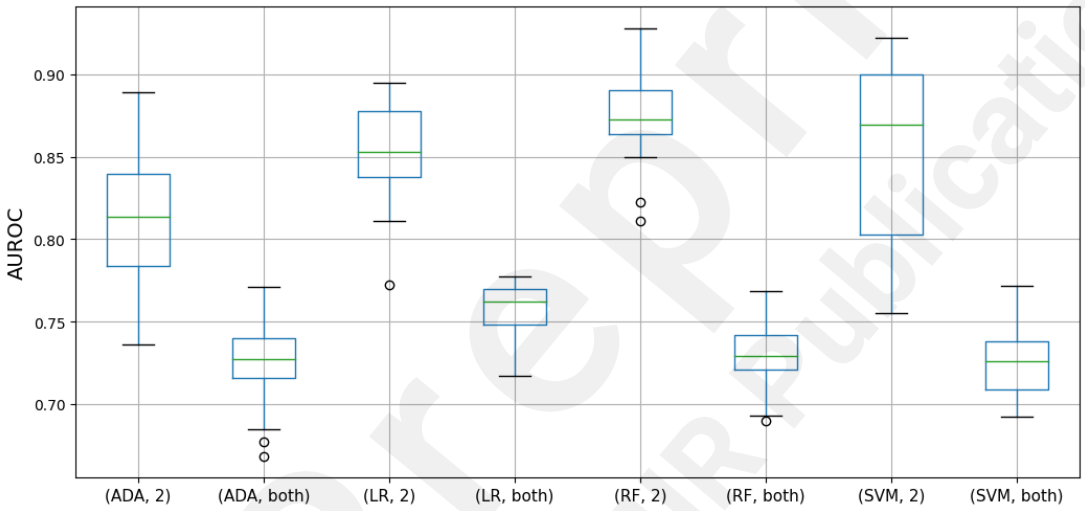


Figure 2. Influence of a pooled test set. Comparison of performance (AUROC) of models, which were developed on the first (Figure 2a) and second (Figure 2b) wave cohort. After model development on a single cohort, the model complexity was fixed. Evaluation was carried out on the data of the first or second wave only and on pooled data. The use of a pooled set results in a significant reduction in performances for all four methods, except of the ADA, which failed in the first wave. (1: first wave data, 2: second wave data, both: pooled data from both waves, SVM: support vector machine, LR: logistic regression, RF: random forest, ADA: AdaBoost.) Box indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.

Testing of models developed within one cohort on the data of the other cohort

Models trained on second wave´s data showed, when applied on first wave´s data, a good prediction for non-survivors but an insufficient classification for survivors. The opposite cross-prediction (training set: first wave, testing set: second wave) showed the inverse behaviour correctly classifying survivors and incorrectly predicting non-survivors. Thus, using a “specialized” model, i.e. developed and trained on patients of the one wave, it was not possible to make sufficient predictions on the

other wave patients and vice versa.

Fig. 3a

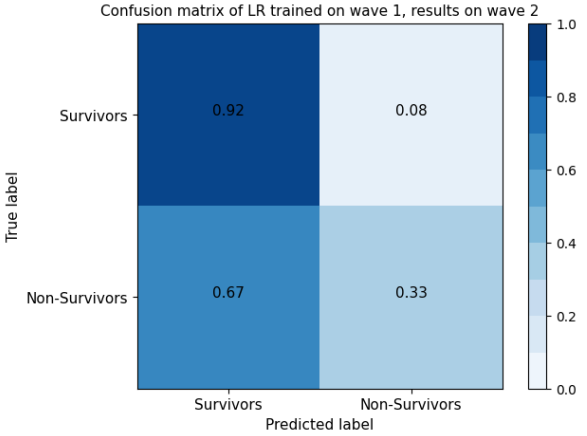


Fig. 3b

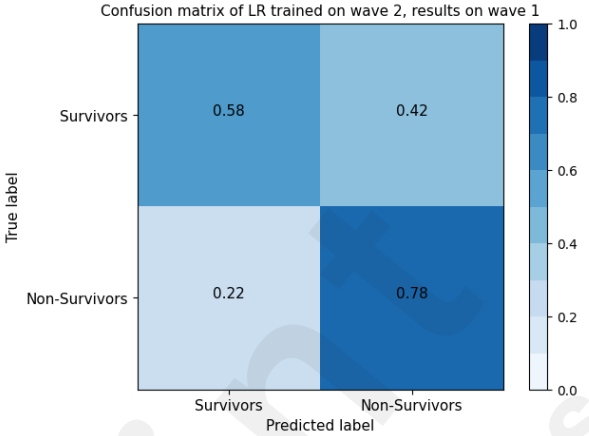


Figure 3. Confusion matrices for cross-prediction. Confusion matrices outlining the results of predicting survival in the second wave cohort with a LR model developed on the first wave cohort (Figure 3a) and in the first wave cohort with a LR model developed on the second wave cohort (Figure 3b). The number inside the cell denotes the ratio of the class that was predicted in relation to the number of instances of an actual class according to the labels on the axes. Numbers are mapped onto colors using a colormap ranging from light blue to dark blue. A darker blue on diagonal cells (true positives and true negatives) and a lighter blue on non-diagonal cells (false positives and false negatives) are considered as better result. (LR: logistic regression.)

Convex hull analysis

We evaluated the intersection of the convex hulls of the first and second wave COVID-19 patients across different attribute spaces in various dimensions in comparison to that of randomly selected patient sets of the same size from the pooled data. We defined the intersection as the ratio of the second wave data points inside the convex hull of the first wave data points. Figure 4 illustrates the intersection range of convex hulls of two waves for 200 randomly chosen attribute spaces with dimensions between 2 and 8. Lower intersection of the convex hulls for the first and second wave populations signals more inhomogeneous distribution when compared to the randomly selected sets of patients. The null hypothesis stating that the mean intersection of convex hulls of two waves is identical to the mean intersection of two sets of randomly selected patients is strongly rejected by the outcome of the t-test, see Table S8 in the Multimedia appendix.

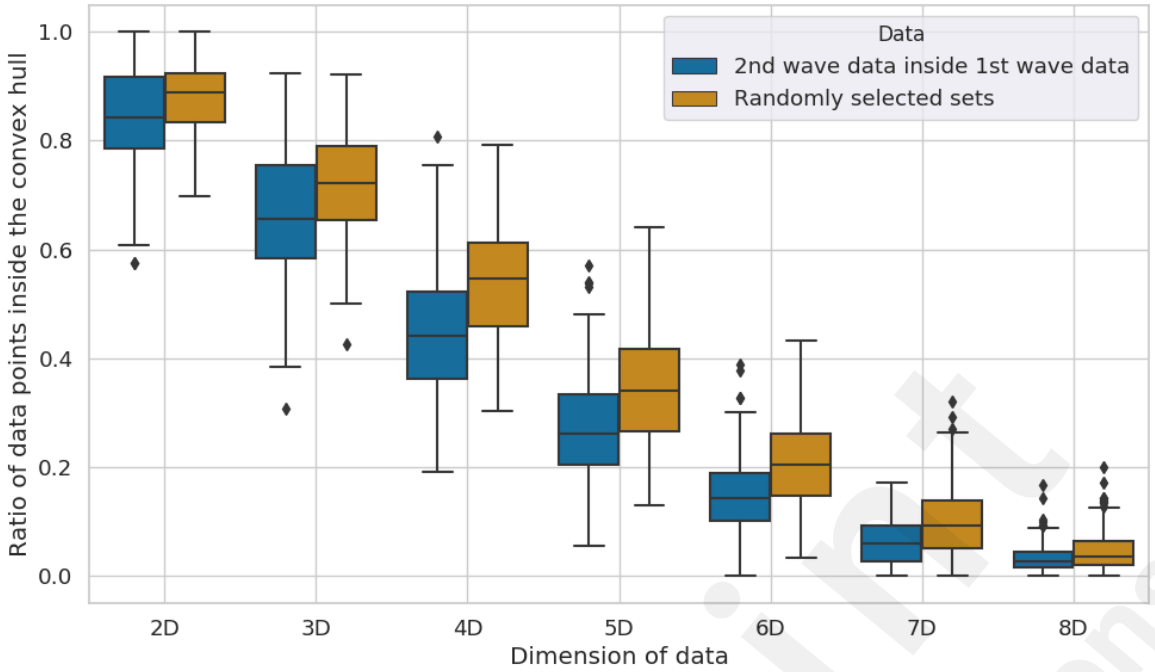
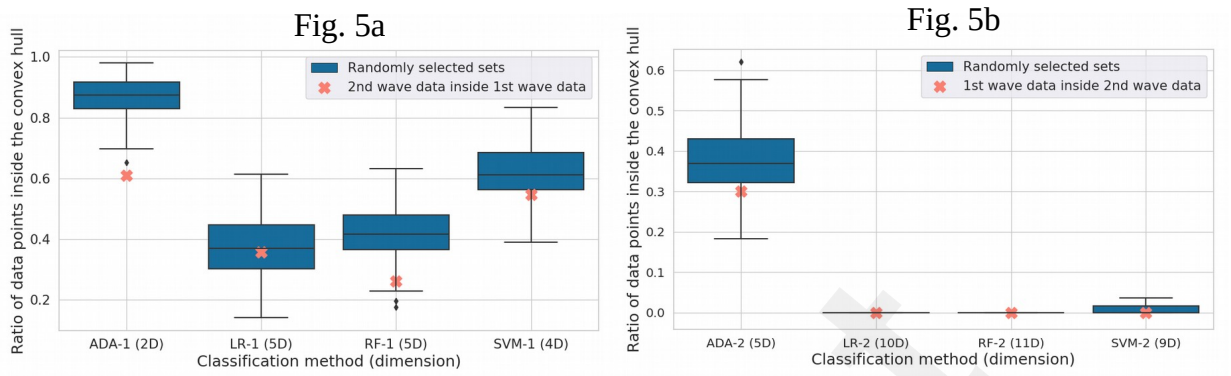


Figure 4. Convex hull analysis of populations separated regarding the two waves and a random pattern population. Box plots show the ratio of the second wave data points inside the convex hull of the first wave data points for 200 different feature spaces with dimensions between 2 to 8. Similar boxplots are depicted for randomly selected patient sets of the same size (63 and 54) from the pooled data. Box indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.

We also analyzed the convex hull coverage across the first and second wave populations for the fixed attribute spaces (complexity) used in the developed consensus models. As depicted in the Figures 5a and 5b, convex hull coverages of the first and second wave populations lie below the convex hull coverage distributions for 200 randomly sampled populations from the pooled data with the same sizes as the first and second wave populations.

Figure 5. Comparison of convex hull coverage for fixed attribute spaces of the consensus



models. The convex hull coverage across the first and second wave populations in comparison with the convex hull coverage of 200 randomly sampled cohorts of patients with the same sizes as the first and the second wave cohorts. The results are depicted for fixed attribute spaces of the consensus models. **a.** The ratio of the second wave data points inside the first wave convex hull. **b.** The ratio of the first wave data points inside the second wave convex hull. The size of patient cohorts are too small for any convex hull coverage in higher dimensions than 9D. Box indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.

Discussion

Model development and performance

For an early phase of the pandemic, we developed a row of models for the prediction of mortality in ventilated ICU patients suffering from COVID 19. Our models contained a small number of parameters (4-5 features), which were already known as predictive from earlier publications. Among the four machine learning methods assessed, the conventional LR model is a standard statistical method, which has been heavily used by critical care professionals and provides easily interpretable results. SVM in contrast to logistic regression, is a non-linear technique allowing to capture higher order interactions between parameters, and therefore to model complex patient physiology. This allows to enhance both prediction and explanatory power [19]. Decision tree classifier (DT) allows physicians to estimate an outcome for a patient based on a usually small number of decisions made one after another, is therefore interpretable, and possesses a high potential to be applied on bedside. However, it is prone to overfitting, as it relies on a particular cut-off value of parameters. In our work, we did not use single DT but rather state-of-the art ensemble methods relying on boosting and bagging techniques, ADA and RF correspondingly. RF comprises an ensemble of decision trees and is intended to reduce generalization errors of single trees. ADA relies on an ensemble of trees, where each subsequent tree tends to focus on harder-to-classify examples [20]. We decided against the use of any types of artificial neural networks (ANNs) in our analysis for several reasons: Firstly, a small sample size reduces capabilities of neural networks and secondly, ANNs are hard to interpret hampering clinical use.

The models reached intermediate to good accuracies with AUROC between 0.58 and 0.82. During the second wave, when the number of patients strongly increased, we tried to add the datasets of the new patients to existing models and expected an even better performance. But surprisingly, the opposite was the case. The addition of the second wave´s patients resulted in relevant decrease of the predictive performance, reducing it to the range from 0.62 to 0.7. Despite almost doubling of the

cohort size, the performance of the models did not increase and even dropped when compared to models developed in a single cohort. Nevertheless, if the patients of the second wave were analysed isolated, it was again possible to produce models with a much better performance than in the first wave, ranging between 0.84 and 0.86. However, these models to large extent used different parameters than the preceding models. Thus, using such a “specialized” model, e.g. trained on patients of the first wave it is not possible to make sufficient predictions on second wave patients and vice versa.

Model development on heterogeneous cohorts

The performance of a selected model is determined by its ability to separate classes on the given data. However, if different classifiers perform poorly on one cohort and better on the other cohort, it reflects how homogeneous the underlying data is, especially regarding differences between survivors and non-survivors. In other words, performance of models indirectly reflects information content in the data. If we compare the performance of all models on the first wave data with the performance on the second wave, we saw that on average all models performed better on the second wave data.

The poor performance in wave one comes along with the strikingly higher ICU length of stay, which was nearly doubled when compared to the second wave. It is obvious that the first wave data represents the first confrontation of physicians with a completely new disease. Thus, knowledge about the disease but also about beneficial therapy strategies was lacking in the first months of the pandemic and needed to be acquired first. These points might have led to longer ICU stays with a higher rate of undesirable events like a secondary pneumonia or sepsis. Although the reasons for these findings are still not fully clear, the sheer duration of ICU treatment might cause a higher inhomogeneity in the first wave cohort. Additionally, other clinical features, like a lower SOFA score or lower required ventilator pressures, could indicate a slightly lower severity in the second wave.

Taken together, the different cohort characteristics but also the observed relevant differences in predictive performance suggest that the resulting pooled population is structurally strongly inhomogeneous. Therefore, while merging the cohorts of two wave the methods struggle to find a separative hyperplane, resulting in a mediocre performance. The finding that models developed within different cohorts would use mostly different parameters also supports the conclusion that data of the two waves could be considered as different populations. From a medical point of view, this is absolutely not surprising taking therapeutic improvements and subsequently a decreasing mortality of COVID-19 patients into account [5, 6, 8].

Usually one can contrast “internal” model performance on structurally similar, previously unseen data gathered from the same hospital used for model training with “external” model performance on new, previously unseen data from different hospital systems. It has been addressed that machine learning models may perform worse in external cohorts due to several reasons, among which are different protocols, confounding variables or heterogeneous populations [3, 21-23]. In our study, the risk of confounding due to different protocols or data acquisition systems was significantly reduced by the fact that data for both cohorts were obtained from the same hospital. Therefore, this again confirms that datasets from different waves of disease represent strongly heterogeneous populations.

The fact that performance of models on pooled data did not increase and even significantly dropped (see Table S4 in Multimedia appendix) confirms that pooling data should be performed with large precaution, even given that analysis is restricted to the same hospital.

Cross-prediction

When we carried out a cross-prediction, surprisingly the models trained on wave one data showed a good accuracy for prediction of survival while the models trained on wave two predicted the opposite, i.e. the non-survival more precisely. At first glance, this finding appears contradictory. However, it must be remembered that the wave 2 patients had lower severity in terms of SOFA score and other clinical parameters. We can conclude that both survivors and non-survivors of the first wave were more severely ill than in the second wave. Therefore, when we apply models developed within the first wave onto the second wave, a majority of patients look like survivors for the model. The opposite holds for application of second wave models - then, patients of the first wave are similar to non-survivors of the second wave.

Convex hull analysis

The points mentioned so far give a clear indication that it is reasonable to consider populations of different waves as independent entities. This evidence can be fully confirmed by the Convex hull analysis, which demonstrates a significant difference between populations. The insufficient intersection of the convex hulls between the first and second wave populations could be considered the primary cause of the observed decrease in the AUROC of the developed consensus models for a population after adding data from the other population.

Future perspectives

It can be assumed that the COVID situation will remain an extremely dynamic one. While first and second waves already showed quite different properties, it is highly likely that even more relevant factors will enter the field, what might change the situation in the future. These are on the one hand mutants of the virus which spread in several parts of the world and exhibit clinical features different from the wild-type [24]. Moreover, the vaccination of the general public proceeds and will lead to changes of the biometric parameters of ICU patients and thus to outcome changes of former highly endangered populations like elderly citizens [25, 26]. As long as the vaccination programs are not completed, it can be expected that the population requiring ICU treatment will shift towards younger, unvaccinated patients. These relevant changes might on the one side reduce the applicability of the models developed on data of earlier cases. On the other side, it makes datasets from a potential next wave even less usable for a pooled population than before.

Limitations

Our study has limitations that need to be considered. Firstly, the sample size of our study was small and secondly, data derived from a single centre. The second drawback, however, turns into an advantage in light of a purpose of our study - to compare how predictive models would perform in temporarily disjoint populations. The monocentric character of our study also ensured that data acquisition and processing were conducted in a similar fashion. Further validation studies including a larger database also with data from multiple hospitals would be very desirable.

Conclusions

The development of COVID-19 related decision support systems is hampered by the lack of a sufficient amount of the especially protected health-related data in many countries. In a more general perspective, our work can serve as a warning and a reminder not to believe that a larger dataset is the solution to all machine learning problems. Rather, it shows that inhomogeneous data used to develop models can lead to serious problems. With the convex hull analysis, we offer a solution for this problem. The outcome of such an analysis can raise concerns if the pooling of different datasets would cause inhomogeneous patterns. If a check for homogeneity is added to the data preparation

step before a model development, the generated models might exhibit a significantly better performance.

Acknowledgements

This publication of the SMITH consortium was supported by the German Federal Ministry of Education and Research, grant numbers 01ZZ1803B and 01ZZ1803M.
We would like to thank for the support, which Moein Einollahzadeh Samadi received as a member of the “Helmholtz School for Data Science in Life Earth and Energy” (HDS-LEE).

Author´s contributions

JB created the dataset of patients and applied for consent of the local ethical review board. KS, MES and AS analyzed the patient data and developed the prediction model. SF gave medical advice during development of the models. SF and JB interpreted the results from a medical perspective. SF, KS, JB and AS wrote the manuscript. All authors read and approved the final manuscript.

Conflicts of Interest

None declared.

Multimedia appendix

Additional information about the development of prediction models and additional tables.

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. 2019 2019/01/01;25(1):44-56. PMID: 30617339. doi: 10.1038/s41591-018-0300-7.
2. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328. PMID: 32265220. doi: 10.1136/bmj.m1328.
3. Goncalves J, Yan L, Zhang H-T, Xiao Y, Wang M, Guo Y, et al. Li Yan et al. reply. Nature Machine Intelligence. 2021 2021/01/01;3(1):28-32. doi: 10.1038/s42256-020-00251-5.
4. Schwab P, Mehrjou A, Parbhoo S, Celi LA, Hetzel J, Hofer M, et al. Real-time prediction of COVID-19 related mortality using electronic health records. Nature communications. 2021 Feb 16;12(1):1058. PMID: 33594046. doi: 10.1038/s41467-020-20816-7.
5. Karagiannidis C, Windisch W, McAuley DF, Welte T, Busse R. Major differences in ICU admissions during the first and second COVID-19 wave in Germany. The Lancet Respiratory medicine. 2021 Mar 5. PMID: 33684356. doi: 10.1016/s2213-2600(21)00101-6.
6. Kluge S, Janssens U, Spinner CD, Pfeifer M, Marx G, Karagiannidis C. Recommendations on Inpatient Treatment of Patients With COVID-19. Dtsch Arztebl International. 2021;118(1-2):1-7. PMID: 33531113.
7. Anesi GL, Jablonski J, Harhay MO, Atkins JH, Bajaj J, Baston C, et al. Characteristics, Outcomes, and Trends of Patients With COVID-19-Related Critical Illness at a Learning Health System in the United States. Annals of internal medicine. 2021 Jan 19. PMID: 33460330. doi: 10.7326/m20-5327.
8. Horwitz LI, Jones SA, Cerfolio RJ, Francois F, Greco J, Rudy B, et al. Trends in COVID-19

Risk-Adjusted Mortality Rates. Journal of hospital medicine. 2021 Feb;16(2):90-2. PMID: 33147129. doi: 10.12788/jhm.3552.
9. Kurtz P, Bastos LSL, Dantas LF, Zampieri FG, Soares M, Hamacher S, et al. Evolving changes in mortality of 13,301 critically ill adult patients with COVID-19 over 8 months. Intensive Care Medicine. 2021 2021/05/01;47(5):538-48. PMID: 33852032. doi: 10.1007/s00134-021-06388-0.
10. Courrieu P. Three algorithms for estimating the domain of validity of feedforward neural networks. Neural Networks. 1994;7(1):169-74.
11. Ostrouchov G, Samatova NF. On FastMap and the convex hull of multivariate data: toward fast and robust dimension reduction. IEEE transactions on pattern analysis and machine intelligence. 2005;27(8):1340-3.
12. Zhou X, Shi Y, editors. Nearest neighbor convex hull classification method for face recognition. International Conference on Computational Science; 2009: Springer.
13. Graham RL. An efficient algorithm for determining the convex hull of a finite planar set. Info Pro Lett. 1972;1:132-3.
14. Jarvis RA. On the identification of the convex hull of a finite set of points in the plane. Information processing letters. 1973;2(1):18-21.
15. Vincent JL, Moreno R, Takala J, Willatts S, De Mendonca A, Bruining H, et al. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. Intensive Care Med. 1996 Jul;22(7):707-10. PMID: 8844239. doi: 10.1007/BF01709751.
16. Lambden S, Laterre PF, Levy MM, Francois B. The SOFA score-development, utility and challenges of accurate assessment in clinical trials. Crit Care. 2019 Nov 27;23(1):374. PMID: 31775846. doi: 10.1186/s13054-019-2663-7.
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
18. Van Rossum G, Drake FL. PYTHON 2.6 Reference Manual. 2009.
19. Johnson AE, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine Learning and Decision Support in Critical Care. Proceedings of the IEEE Institute of Electrical and Electronics Engineers. 2016 Feb;104(2):444-66. PMID: 27765959. doi: 10.1109/jproc.2015.2501978.
20. Wyner AJ, Olson M, Bleich J, Mease D. Explaining the success of adaboost and random forests as interpolating classifiers. The Journal of Machine Learning Research. 2017;18(1):1558-90.
21. Cabitza F, Rasoini R, Gensini GF. Unintended consequences of machine learning in medicine. Jama. 2017;318(6):517-8. PMID: 28727867.
22. Mårtensson G, Ferreira D, Granberg T, Cavallin L, Oppedal K, Padovani A, et al. The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. Medical Image Analysis. 2020;66:101714. PMID: 33007638.
23. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLOS Medicine. 2018;15(11):e1002683. PMID: 30399157. doi: 10.1371/journal.pmed.1002683.
24. Harvey WT, Carabelli AM, Jackson B, Gupta RK, Thomson EC, Harrison EM, et al. SARS-CoV-2 variants, spike mutations and immune escape. Nature Reviews Microbiology. 2021 2021/07/01;19(7):409-24. PMID: 34075212. doi: 10.1038/s41579-021-00573-0.
25. Haas EJ, Angulo FJ, McLaughlin JM, Anis E, Singer SR, Khan F, et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. The Lancet. 2021 2021/05/15/;397(10287):1819-29. PMID: 33964222. doi: https://doi.org/10.1016/S0140-6736(21)00947-8.

26. Vasileiou E, Simpson CR, Shi T, Kerr S, Agrawal U, Akbari A, et al. Interim findings from first-dose mass COVID-19 vaccination roll-out and COVID-19 hospital admissions in Scotland: a national prospective cohort study. The Lancet. 2021 2021/05/01/;397(10285):1646-57. PMID: 33901420. doi: https://doi.org/10.1016/S0140-6736(21)00677-2.

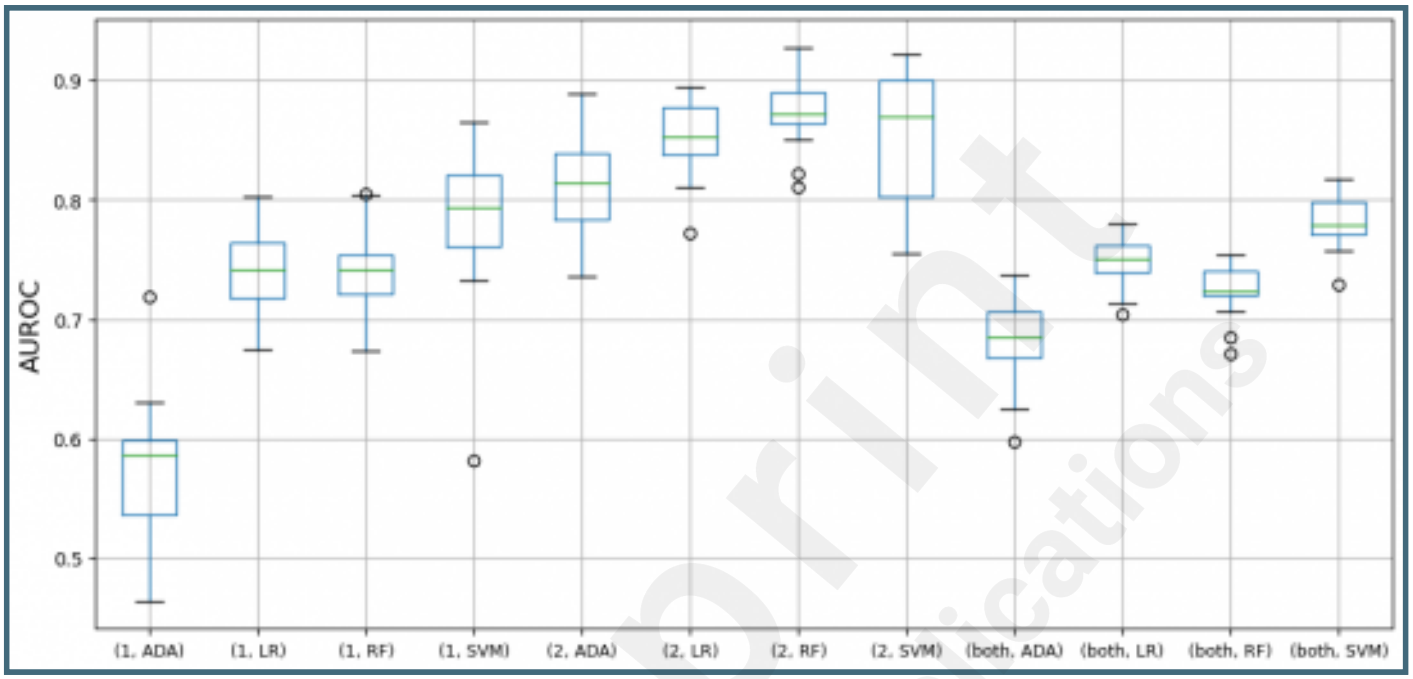
Abbreviations

ADA (Boost)	Adaptive boosting
AI	Artificial Intelligence
COVID 19	Coronavirus Disease 2019
DT	Decision tree
ICU	Intensive care unit
LR	Logistic regression
MV	Mechanical ventilation
p_aCO_2	Arterial partial pressure of CO_2
PEEP	Positive end-expiratory pressure
P EI	End-inspiratory pressure
RF	Random forest
SOFA	Sequential organ failure assessment
SVM	Support vector machine

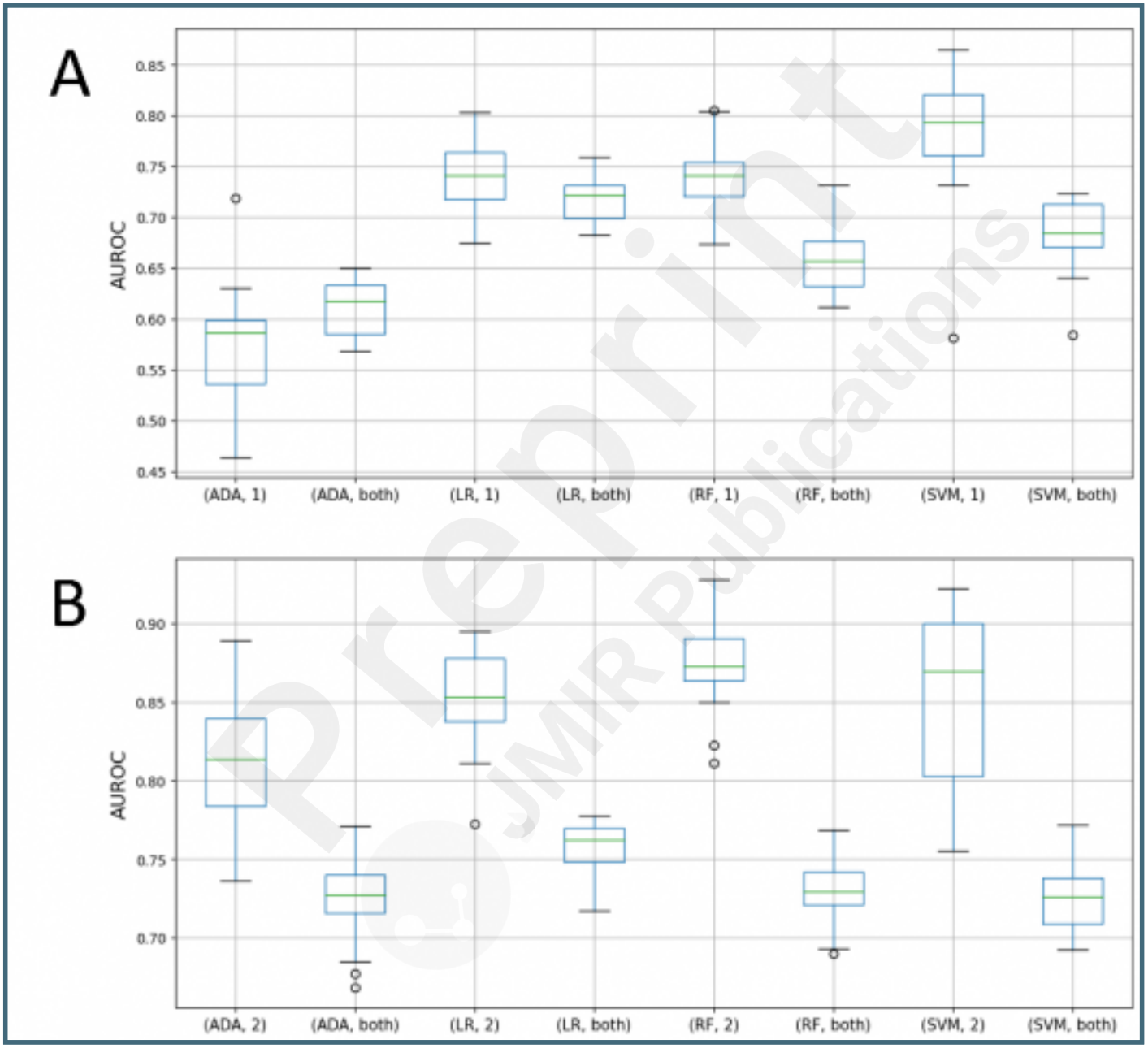
Supplementary Files

Figures

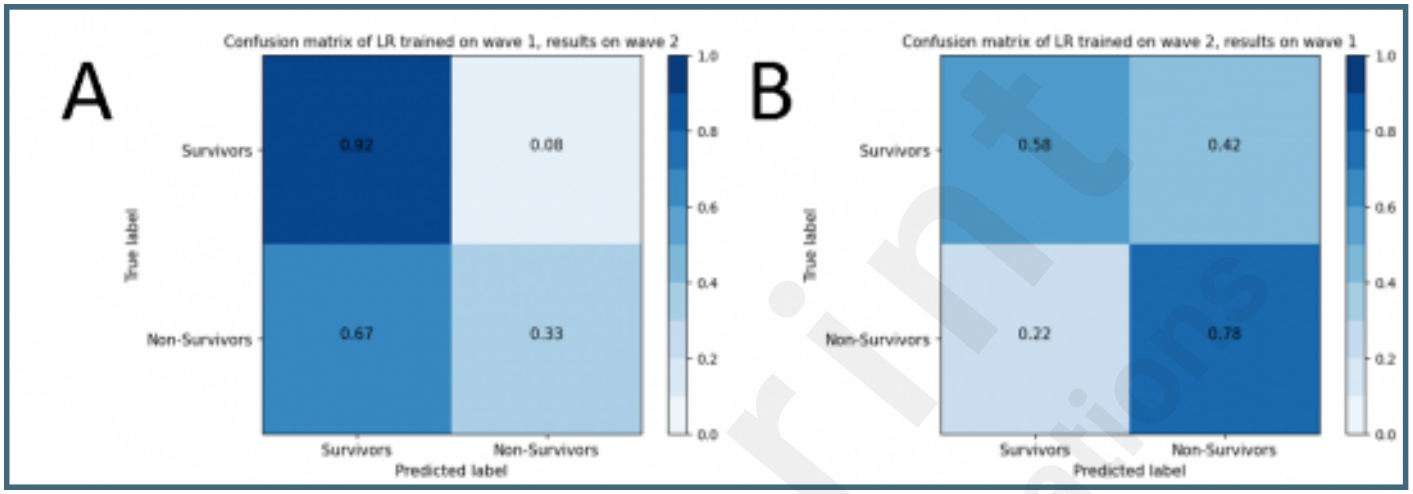
Performance of developed models with fixed complexity. Performance is depicted using AUROC in the nested CV procedure. From left to right: four boxplots show performances of the models developed and tested with data from the first wave only, the second wave only and the pooled data from both waves, respectively. (1: first wave data, 2: second wave data, both: pooled data, SVM: support vector machine, LR: logistic regression, RF: random forest, ADA: AdaBoost.) Boxes indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.



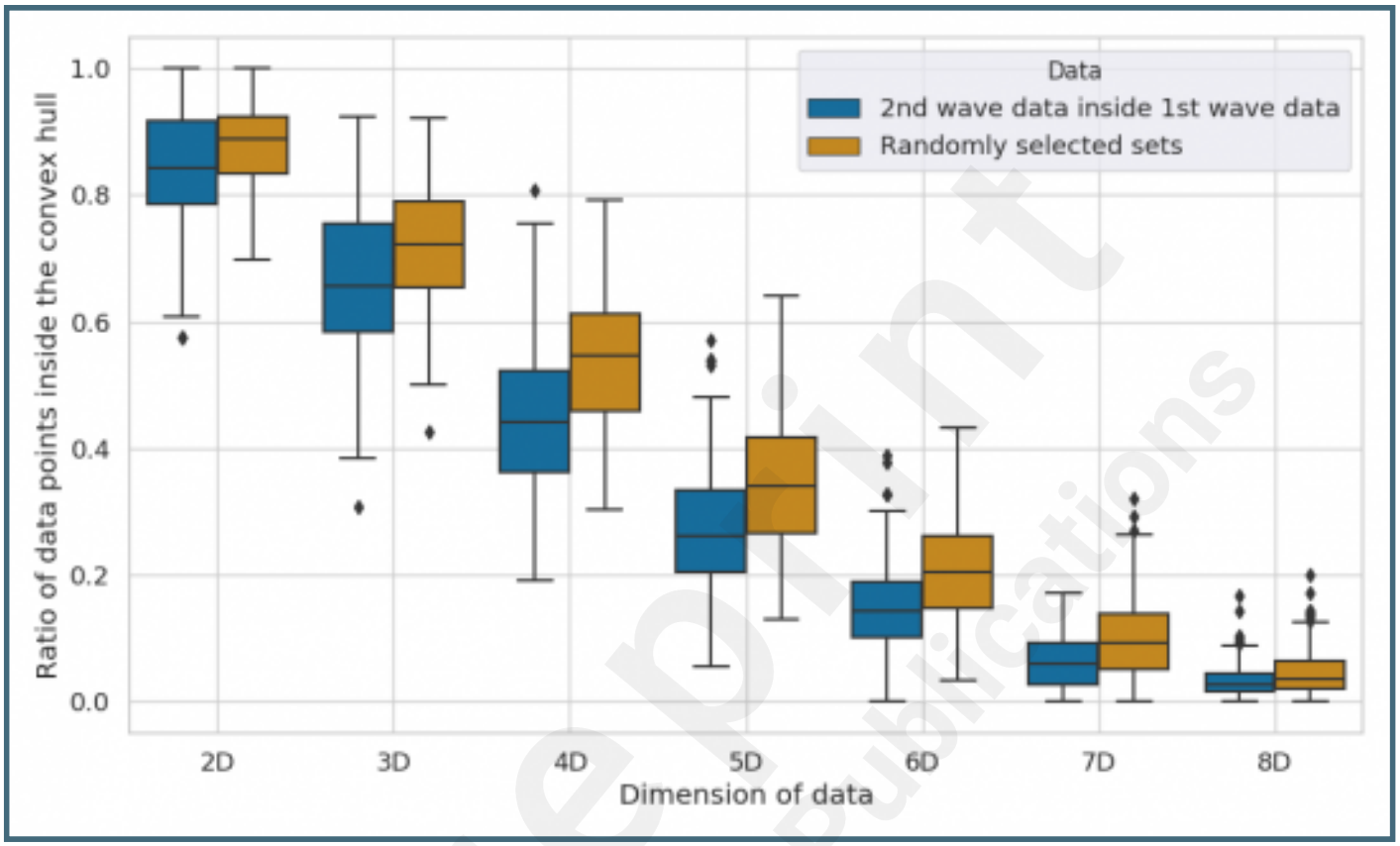
Influence of a pooled test set. Comparison of performance (AUROC) of models, which were developed on the first (A) and second (B) wave cohort. After model development on a single cohort, the model complexity was fixed. Evaluation was carried out on the data of the first or second wave only and on pooled data. The use of a pooled set results in a significant reduction in performances for all four methods, except of the ADA, which failed in the first wave. (1: first wave data, 2: second wave data, both: pooled data from both waves, SVM: support vector machine, LR: logistic regression, RF: random forest, ADA: AdaBoost.) Box indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.



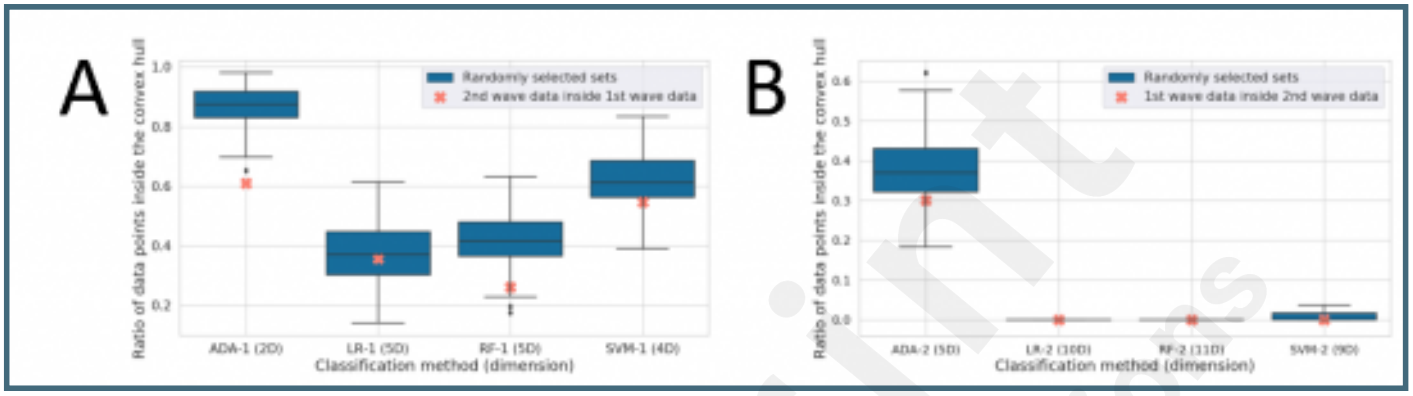
Confusion matrices for cross-prediction. Confusion matrices outlining the results of predicting survival in the second wave cohort with a LR model developed on the first wave cohort (A) and in the first wave cohort with a LR model developed on the second wave cohort (B). The number inside the cell denotes the ratio of the class that was predicted in relation to the number of instances of an actual class according to the labels on the axes. Numbers are mapped onto colors using a colormap ranging from light blue to dark blue. A darker blue on diagonal cells (true positives and true negatives) and a lighter blue on non-diagonal cells (false positives and false negatives) are considered as better result. (LR: logistic regression).



Convex hull analysis of populations separated regarding the two waves and a random pattern population. Box plots show the ratio of the second wave data points inside the convex hull of the first wave data points for 200 different feature spaces with dimensions between 2 to 8. Similar boxplots are depicted for randomly selected patient sets of the same size (63 and 54) from the pooled data. Box indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.



Comparison of convex hull coverage for fixed attribute spaces of the consensus models. The convex hull coverage across the first and second wave populations in comparison with the convex hull coverage of 200 randomly sampled cohorts of patients with the same sizes as the first and the second wave cohorts. The results are depicted for fixed attribute spaces of the consensus models. A. The ratio of the second wave data points inside the first wave convex hull. B. The ratio of the first wave data points inside the second wave convex hull. The size of patient cohorts are too small for any convex hull coverage in higher dimensions than 9D. Box indicating the IQR and the median, the whiskers indicating the 1.5x IQR and the dots indicating outliers.



Multimedia Appendixes

Additional information about the development of prediction models and additional tables.
URL: http://asset.jmir.pub/assets/d58430e832e60164b9616ab006dc0109.docx