Journal of Medical Systems (2022) 46: 23
https://doi.org/10.1007/s10916-022-01807-1

CLINICAL SYSTEMS

Domain Shifts in Machine Learning Based Covid-19 Diagnosis From Blood Tests

Theresa Roland[1] · Carl Böck[2] · Thomas Tschoellitsch[2] · Alexander Maletzky[3] · Sepp Hochreiter[1] · Jens Meier[2] · Günter Klambauer[1]

Received: 1 October 2021 / Accepted: 10 February 2022 / Published online: 29 March 2022

Abstract

Many previous studies claim to have developed machine learning models that diagnose COVID-19 from blood tests. However, we hypothesize that changes in the underlying distribution of the data, so called domain shifts, affect the predictive performance and reliability and are a reason for the failure of such machine learning models in clinical application. Domain shifts can be caused, e.g., by changes in the disease prevalence (spreading or tested population), by refined RT-PCR testing procedures (way of taking samples, laboratory procedures), or by virus mutations. Therefore, machine learning models for diagnosing COVID-19 or other diseases may not be reliable and degrade in performance over time. We investigate whether domain shifts are present in COVID-19 datasets and how they affect machine learning methods. We further set out to estimate the mortality risk based on routinely acquired blood tests in a hospital setting throughout pandemics and under domain shifts. We reveal domain shifts by evaluating the models on a large-scale dataset with different assessment strategies, such as temporal validation. We present the novel finding that domain shifts strongly affect machine learning models for COVID-19 diagnosis and deteriorate their predictive performance and credibility. Therefore, frequent re-training and re-assessment are indispensable for robust models enabling clinical utility.

Keywords Machine learning · Domain shift · COVID-19 · Blood test

Introduction

Reverse transcription polymerase chain reaction (RT-PCR) [1] remains the gold standard test for the coronavirus disease 2019 (COVID-19) [2]. However, RT-PCR tests are expensive, time-consuming, and not suited for high-throughput or large-scale testing efforts. In contrast, antigen tests [3] are cheap and fast, but they come with considerably lower sensitivity than RT-PCR tests [4]. Instead of RT-PCR tests or antigen tests, routine blood tests can be automatically scanned for COVID-19: machine learning (ML) models can predict the diagnoses on the basis of blood tests, which are taken in the routine processes of the hospital. The routine blood tests are acquired anyway, therefore, no additional efforts are caused by screening with ML models. Routine screening of the blood tests would allow frequent, fast and broad testing at low cost, thus providing a powerful tool to reduce new outbreaks in the hospital [5, 6]. Especially in developing countries with limited testing capacities, the ML enhanced tests can evolve into an efficient tool in combating a pandemic.

ML methods offer very different ways to help confining the spread of infectious diseases [7–13], e.g., in developing vaccines and drugs for the treatment of COVID-19 [14–16]. COVID-19 diagnosis and the patient's prognosis can be predicted from chest CT-scans, X-rays [17–25] or sound recordings of coughs or breathing [26–28]. Furthermore, it has been shown that ML models based on blood tests are capable of detecting COVID-19 infection [29–43]. Other outcomes, such as survival or admission to an intensive care unit can be predicted based on cheap and fast tests, such as blood tests [44–52].

Topical Collection on *Clinical Systems*

✉ Theresa Roland
theresa.roland@jku.at

[1] ELLIS Unit Linz, LIT AI Lab, Institute for Machine Learning, Johannes Kepler University Linz, Linz, Austria

[2] Department of Anesthesiology and Critical Care Medicine, Kepler University Hospital GmbH, Johannes Kepler University Linz, Linz, Austria

[3] RISC Software GmbH, Hagenberg i.M., Austria

In this study, we first reveal the *presence of domain shifts* in COVID-19-related blood test datasets. Second, we evaluate the ML models for prediction of COVID-19 diagnosis and mortality risk with different assessment strategies to demonstrate that these *domain shifts diminish the predictive performance*. Third, we compare the expected and actual performance to show how *model credibility is decreased by domain shifts*.

Domain Shifts

Good generalization of ML models is only possible if the training data and future (test) data arise from the same underlying distribution. Deviations between training and test data distribution are a well known challenge in medical [53] and biological systems and in other real-world applications [54]. The failure of generalization on the test set and limited reliability of ML models in clinical settings has already been discussed in literature [55]. The negative effects and the necessity for countering these domain shifts in various complex biological systems have to be considered for ML models [56]. The necessity for critical appraisal and reporting of models for diagnosis and prognosis has been published in the context of the TRIPOD-AI guideline [57].

The same underlying distribution of training and future data also cannot be guaranteed during pandemics. Examples of potential domain shifts in COVID-19 related datasets are plotted in Fig. 1. Most of the previous COVID-19 ML studies evaluated their models by cross-validation, bootstrapping or fixed splits on randomly drawn samples [29–33, 37–43], which disregard changes in the underlying distribution over time, so-called domain shifts.

The domain shifts [54, 59, 60] can occur because of changes of the probability of observing a certain RT-PCR test result, which strongly changes during the pandemic. It can also change with the distribution of the blood test features, which are also affected by the overall pandemic course, but also, e.g., with the time of the year without connection to the pandemic [61]. The joint distribution of patient features and labels can change, e.g., with new virus mutations. Machine learning and statistical approaches model the probability to observe a certain RT-PCR test result given a patient. However, the RT-PCR test results might also be affected by changing test technologies or changing thresholds.

Neglecting and insufficiently countering these domain shifts can lead to undesired consequences and failures of the models. The domain shifts can lead to *degrading of predictive performance over time*, because standard ML approaches are unable to cope with domain shifts over time [54, 59, 60]. Further, the domain shifts can cause *unreliable performance estimates*. These performance estimates might be overoptimistic and can deviate significantly from the actual performance [62].

The ML models in our experiments do not require additional expensive features [32–34, 45–52]. The RT-PCR test results serve as the ground truth for the COVID-19 diagnosis (positive or negative) prediction. The in-hospital death is the label for the mortality (survivor or deceased) prediction of COVID-19 positive patients. The models are trained and evaluated on a large-scale dataset, which exceeds the dataset size of many small-scale studies [29–33, 43–46, 52] by far.

The findings of our work do not only apply to COVID-19 datasets, but also to future pandemics, other medical datasets and even to datasets from other fields, where domain shifts might play a role.

Materials and Methods

Ethics approval for this study was obtained from the ethics committee of the Johannes Kepler University, Linz (approval number: 1104/2020). In our study, we analyze anonymized data only. The dataset was collected, pre-processed and the blood tests were merged with the RT-PCR tests.

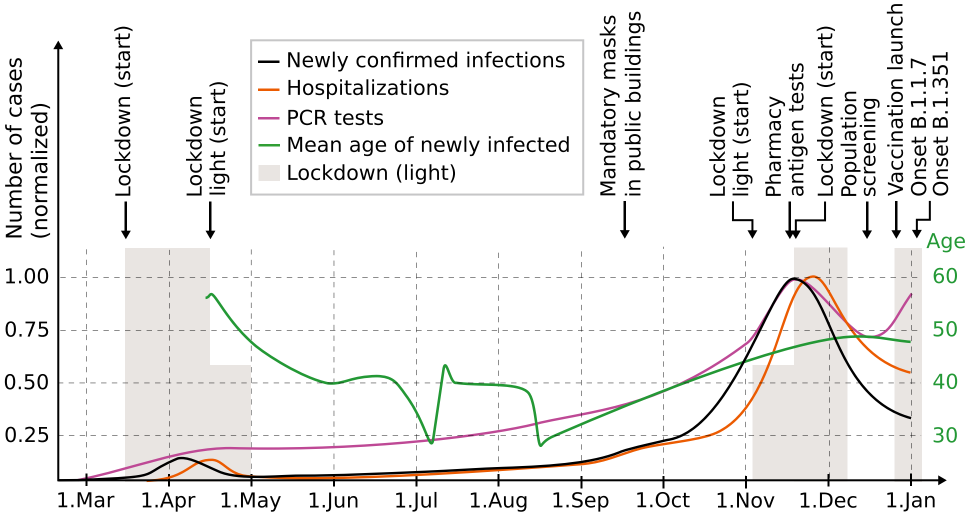

Fig. 1 Examples of temporal domain shifts in COVID-19 datasets, which might diminish the ML model's predictive performance over time. COVID-19 numbers in Austria over time, illustrating factors causing a temporal domain shift. The numbers are sketched according to data from the Austrian BMSGPK [58]

As a first step, we plotted the statistics of the blood test parameters over time to visualize fluctuations of the statistics indicating the presence of domain shifts. To answer, whether domain shifts in the dataset cause degrading of predictive performance, we implemented different assessment strategies. To analyze the model credibility, a comparison of expected and actual performance was implemented and examined. Additional experiments and results are presented in the Supplementary Information.

Dataset

The study is conducted on the dataset (Table 1) of the Kepler University Hospital, Med Campus III, Linz, Austria. The nature of the dataset corresponds neither perfectly to a cross-sectional study, since samples are taken at many different time-points, nor to a longitudinal study, since at each time-point a different set of samples is analyzed. Our analyses are based on blood tests, which are acquired in the routine process of the hospital. The features age, sex and hospital admission type (inpatient or outpatient) are added to the samples. If parameters in the blood tests are measured more than once, the most recent one is selected (Fig. 2). In case no COVID-19 test follows the blood test within 48 h in the *2020 cohort*, the blood test samples are discarded. Hence, the *2020 cohort* is biased towards patients, who might already be suspect for being COVID-19 positive and therefore are tested. Additionally, all samples with a deviating RT-PCR test result within the next 48 h are discarded, as the label might be incorrect.

Additionally, we incorporate pre-pandemic blood tests from the year 2019 as negatives to our dataset to cover a wide variety of COVID-19 negative blood tests (*2019 cohort*). The *2019 cohort* does not contain COVID-19 tests, therefore, blood tests with a temporal distance of less than 48 h are aggregated. A temporal distance of 48 h is selected such that the *2019 cohort* resembles the *2020 cohort*. The samples with less than 15 features are dropped from the dataset, all other available blood tests from the year 2019 are incorporated in the dataset. We assume that all patients in the year 2019 have been COVID-19 negative, because the virus has not been detected in Austria at this time. With a large, diverse dataset, the data distribution of the COVID-19 negative samples is broadly covered and learnt by the ML model. The distribution of the negative samples provided to the model during training has to be similar to the test data distribution for high predictive performance. During deployment, the model will be confronted with negative blood tests from a broad spectrum of different health scenarios, therefore, the *2019 cohort* is incorporated during training.

Before the selection of the 100 most frequent features, we include all available blood test parameters from the Med Campus III in Linz. This ranges from standard blood test parameters, such as leucocyte count up to blood tests for rare tropical diseases. Only the COVID-19 antibody tests are discarded from the dataset, as these might be directly related to the COVID-19 status. For the prediction of the COVID-19 diagnosis, the 100 most frequent features in the *2019 cohort* are selected as the feature set. For the mortality task these 100 most frequent features are selected based on the *positives cohort*. The number of measurements for each blood test parameter in the hospital is determined. The blood test parameters, which have been measured most frequently, are selected as input features for the ML models. Each sample requires a minimum of 15 features (minimum of any twelve

Table 1 Dataset with summary of patient characteristics

	N cases[a]	N positives	N negatives	Age (mean ± sd)	Sex (f/m), (f%)	Adm. type (i/o), (i%)[b]
Full dataset (*2019* and *2020 cohort*)	79 884	1037	79 053	53.4 ± 25.3	41 589/38 295 (52.1%)	50 727/29 157 (63.5%)
2019 cohort (pre-pandemic)	70 870	-	70 870	52.8 ± 25.1	36 934/33 936 (52.1%)	42 791/28 079 (60.4%)
2020 cohort (pandemic)	9014	1037	8183	58.0 ± 26.4	4655/4359 (51.6%)	7936/1078 (88.0%)
Negatives cohort	79 053	-	79 053	53.3 ± 25.4	41 213/37,840 (52.1%)	50 020/29,033 (63.3%)
Positives cohort	1037	1037	-	64.3 ± 20.2	455/582 (43.9%)	908/129 (87.6%)
Survivors (with COVID-19)	919	919	-	62.7 ± 20.5	417/502 (45.4%)	790/129 (86.0%)
Deceased (with COVID-19)	118	118	-	76.6 ± 11.8	38/80 (32.2%)	118/0 (100%)
March-October 2020 (training and validation cohort for prospective assessment)	6504	291	6277	57.0 ± 27.3	3416/3088 (52.5%)	5720/784 (87.9%)
November–December 2020 (test cohort for prospective assessment)	2636	785	1982	60.8 ± 24.1	1293/1343 (49.1%)	2335/301 (88.6%)

[a] Multiple samples can be obtained from one case. Therefore, one case can be contained in both, the *positives* and the *negatives cohort*, due to a change of the COVID-19 diagnoses, e.g., the patient might have been infected during the hospital stay, or the patient's coronavirus load might have decreased, yielding a negative test result

[b] Adm. type: Admission type, i: inpatient, o: outpatient

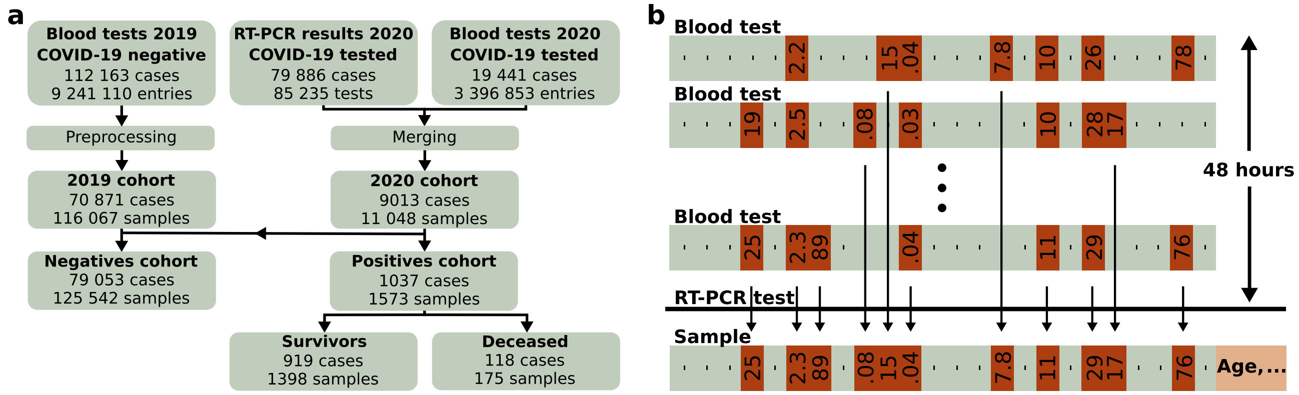

Fig. 2 Large-scale COVID-19 dataset. **a**: A block diagram of the structure of the dataset. The blood tests from 2019 (blood tests 2019) are all negatives and are pre-processed to the *2019 cohort*. The COVID-19 RT-PCR test results and the blood tests are merged to the *2020 cohort*. The *negatives cohort* results from the *2019 cohort* (pre-pandemic samples) and the negative samples of the *2020 cohort*. The positive tested cases (*positives cohort*) are further divided to the cohort with the *survivors* and *deceased*. Note that one case can be in the *negatives* and *positives cohort* due to a change of the COVID-19 status. Multiple samples are obtained from one case, if RT-PCR and blood tests are measured repeatedly. **b**: Aggregation of the blood tests for the COVID-19 tested patients. The blood tests of the last 48 h before the COVID-19 test are merged to one sample. In case a feature is measured multiple times, the most recent one is inserted in the sample. Patient specific data, namely age, sex and hospital admission type, are added to the sample

blood test features and age, sex and hospital admission type). All other features and samples are discarded. Besides the measured blood test values, the selection of the acquired blood test parameters might also contain relevant information. Therefore, for each sample 100 additional binary entries are created, which indicate whether each of the features is missing or measured. The missing values are filled by median imputation. Hence, the models can be applied to blood tests with few measured values. In the full dataset (*2019* and *2020 cohort*) 58.0% and in the *positives cohort* 49.6% of the selected features are missing.

Domain shifts are changes of the distribution over time, therefore, the mean, median and standard deviation, the first and third quantile of exemplary blood test features of the *positives cohort* are displayed in Fig. 3. Indeed, the statistics change over time, which indicate the presence of domain shifts. These eight features are the most frequently measured blood test features in the *positives cohort*.

Machine Learning Methods and Model Selection

We investigate the capability of different ML model classes to predict the COVID-19 diagnoses and the mortality risk. To this end, the predictive performance of self-normalizing neural networks (SNN) [63], K-nearest neighbor (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGB) are compared against each other. The pre-processing, training and evaluation is implemented in Python 3.8.3. In particular, the model classes RF, KNN and SVM are trained with the scikit-learn package 0.22.1. XGB is trained with the XGBClassifier from the Python package XGBoost 1.3.1. The SNN and LR are trained with Pytorch 1.5.0.

The hyperparameters are selected via grid-search on a validation set or via nested cross-validation to avoid a hyperparameter selection bias (Table S2). The training, validation and test splits are conducted on patient level, such that one patient only occurs in one of the sets and the dataset is Z-score normalized based on the mean and standard deviation of the training set.

The models are selected and evaluated based on the area under the receiver operating characteristic curve (ROC AUC) [64], which is a measure of the model's discriminating power between the two classes and is in this case equivalent to the concordance-statistic (c-statistic) for binary outcomes [64]. Further, we report the area under the precision recall curve (PR AUC) [65] and we also calculate threshold-dependent metrics, where the classes are separated into positives and negatives, instead of probability estimates. These metrics are negative predictive value (NPV), positive predictive value (PPV), balanced accuracy (BACC), accuracy (ACC), sensitivity, specificity and the F1-score (F1) [66]. We additionally report the thresholds, which are determined on the validation set to achieve the intended NPV.

Experiments for Model Performance under Domain Shift

In this section, we evaluate whether domain shifts diminish the predictive performance of ML models. A flow chart about the assessments is shown in the supplementary information (Fig. S1). Therefore, five modeling experiments with

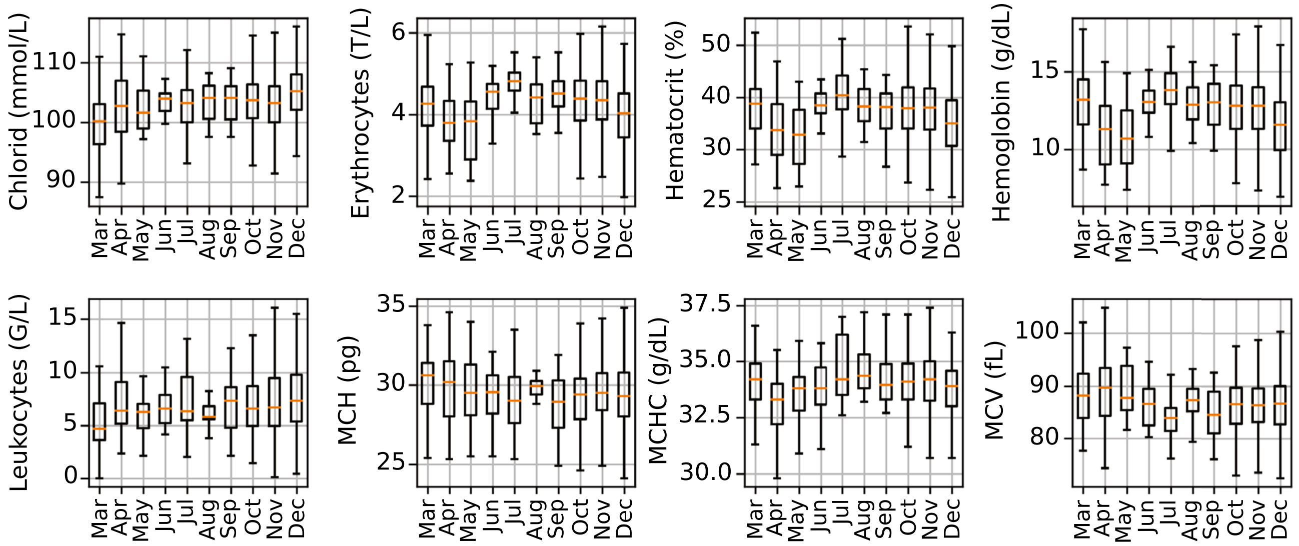

Fig. 3 Statistics of blood test features of the *positives cohort*. The change of the statistics over time indicate a change of the underlying distribution and the presence of domain shifts. Abbreviations: mean cell hemoglobin (MCH), mean corpuscular hemoglobin concentration (MCHC), mean corpuscular volume (MCV)

two prediction tasks and different assessment strategies are set up:

COVID-19 Diagnosis Prediction

i. assessed by random validation with pre-pandemic negatives.

All patients are randomly shuffled and split regardless of the patient cohorts (60% training, 20% validation, 20% testing). Domain shifts are not considered in this experiment. This experiment is performed to obtain an estimate of the predictive performance if there were no domain shifts in the data. This also corresponds to the performance estimates provided in other studies [29–34, 37–43], which we hypothesize to be over-optimistic.

ii. assessed by random validation with recent negatives.

The training and validation sets include the *2019 cohort* and 80% (60% training, 20% validation) of the *2020 cohort*. The test set comprises the remaining samples (20%) of the *2020 cohort*. Therefore, the performance is estimated on patients, who actually were tested for COVID-19. Domain shifts between the *2019 cohort* and the *2020 cohort* are considered. Domain shifts within the *2020 cohort* are not considered. This experiment is executed in order to reveal the effects of biases and domain shifts between the *2019* and *2020 cohort*.

iii. assessed by temporal validation.

The training and validation sets include the *2019 cohort* and the *2020 cohort* before November (80% training, 20% validation). A prospective performance estimation is conducted for the test set with all samples from November and December 2020. By the temporal split, domain shifts over time are considered. In this experiment, it is investigated how the models would perform in real-world environment, where models can only be trained with data from the past and deployed on future data.

Mortality Prediction

iv. assessed by nested cross-validation.

The training (60%), validation (20%) and test (20%) sets comprise the *positives cohort*, which are the positive cases from the *2020 cohort*. Due to the limited number of samples, predictive performance is estimated with five-fold nested cross validation. This experiment is conducted to show the performance estimates, when domain shifts over time within the *positives cohort* are not considered. We hypothesize, that these results, which correspond to the performance estimates in other studies [46–48], are over-optimistic.

v. assessed by temporal validation.

The training and validation sets include the positive cases from 2020 before November (80% training, 20% validation). The test set comprises the cases from November and December. In this experiment, domain shifts over time are considered. In this experiment, by temporal validation, the performance of the models with consideration of the domain shifts is estimated.

The performance estimates obtained by these different assessment strategies are compared. If the underlying distribution of the data remains similar over time, the performance estimates by random cross-validation and temporal cross-validation must also be similar. If the performance estimates of (ii) are different from (i), then former and more recent negatives follow different distributions and the ML models are affected by the domain shifts. If performance estimates from (iii) are lower than those of (i) and (ii), the distribution of the data changes over time, hence indicating the presence and diminishing effects of domain shifts on predictive performance. Equally, changing performance estimates from (iv) to (v) indicate a domain shift over time. The binomial test [67] is used to check, whether the ML model's (SNN, KNN, LR, SVM, RF, XGB) performance estimates in experiment (i) are equal to the estimates in experiment (ii). Similarly, we compare experiment (ii) with (iii) and (iv) with (v).

Experiments for Model Credibility under Domain Shifts

In this experiment, we test whether domain shifts cause deviations of expected and actual performance. The predictive performance would remain similar without domain shifts, but in the presence of domain shifts, the performance could be significantly different and thus domain shifts may be exposed. If the expected and actual performance are different, the diminishing effect of domain shifts on model credibility are revealed.

In this experiment, a standard ML approach is simulated in which a model is trained on data collected in a particular time-period (model training), then assessed on a hold-out set (expected performance) and then deployed (actual performance) (Fig. 4). For example, the deployment in December 2020 is simulated in the following way: First, an XGB model is trained (with the selected hyperparameters of experiment (iii)) on data from July 2019 until October 2020. The expected performance is then determined on data of November 2020. Then the actual performance of the model is evaluated on the subsequent month (December 2020). In other words, the ROC AUC metrics of two subsequent months are compared. The expected performance is determined with a temporal split, which might already be more credible than an expected performance assessed by random cross-validation. The 95% confidence intervals are determined via bootstrapping by sampling 1000 times with replacement.

Results

Model Performance under Domain Shifts

In general, ML models are capable of diagnosing COVID-19 and predicting mortality risk with high ROC AUC values. XGB and RF outperform other model classes in the COVID-19 diagnosis and in the mortality prediction. The comparison of evaluations on different cohorts expose domain shifts and their diminishing effect on predictive performance. Results are reported in terms of threshold-independent performance metrics for the comparison of the models (Tables 2 and 3) as well as threshold-dependent metrics (Tables S3, S4, S5, S6 and S7).

COVID-19 Diagnosis Prediction

i. assessed by random cross-validation with pre-pandemic negatives.

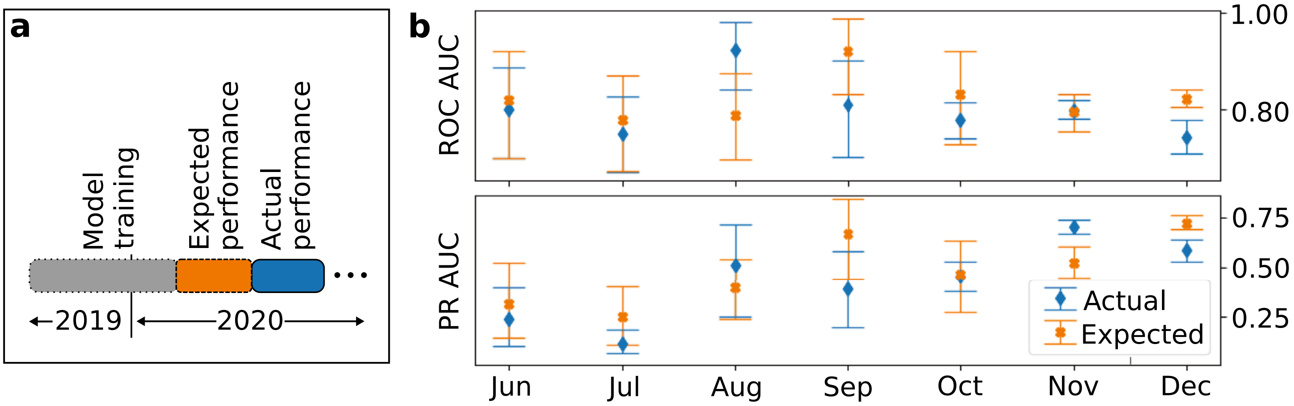

Fig. 4 Comparison of expected and actual performance. **a**: The actual model performance is calculated for each month from June to December 2020 and the expected model performance is calculated on the respective previous month. The ROC AUCs of two subsequent months are compared, which correspond to expected and actual performance. **b**: The expected and actual performance with 95% confidence intervals. The expected and actual ROC AUC is significantly different in December and PR AUC differs significantly in November and December, showing the effect of the domain shifts on model credibility. Note that the PR AUC is sensitive to changes of prevalence

Table 2 Performance metrics of threshold-independent metrics for COVID-19 diagnosis prediction (experiment (i)-(iii)). The mean and the standard deviation (±) for the ROC AUC and PR AUC for the five random seeds are listed. Note that the PR AUC is dependent on the class prior, which changes with the different assessment strategies. E.g., the class prior in the test set in experiment (iii) is higher, because disease prevalence in the evaluation months November and December is higher. The performance estimates of a random estimator (RE) and the best feature (BF) are listed for comparison. The highest performance metrics per experiment are printed in bold

Model	**Experiment (i)**		**Experiment (ii)**		**Experiment (iii)**	
	ROC AUC	PR AUC	ROC AUC	PR AUC	ROC AUC	PR AUC
RE	0.5000 ± 0.0000	0.0124 ± 0.0000	0.5000 ± 0.0000	0.0822 ± 0.0000	0.5000 ± 0.0000	0.3162 ± 0.0000
BF	0.6745 ± 0.0000	0.0221 ± 0.0000	0.6774 ± 0.0000	0.3141 ± 0.0000	0.6623 ± 0.0000	0.5716 ± 0.0000
SNN	0.9567 ± 0.0025	0.4349 ± 0.0306	0.8998 ± 0.0044	0.5577 ± 0.0074	0.7836 ± 0.0053	0.6620 ± 0.0082
KNN	0.9071 ± 0.0000	0.3137 ± 0.0000	0.8432 ± 0.0000	0.4486 ± 0.0000	0.7209 ± 0.0000	0.5712 ± 0.0000
LR	0.9600 ± 0.0008	0.4126 ± 0.0145	0.8878 ± 0.0022	0.4770 ± 0.0086	0.7732 ± 0.0008	0.6467 ± 0.0059
SVM	0.9611 ± 0.0000	0.4268 ± 0.0000	0.9045 ± 0.0000	0.5573 ± 0.0000	0.7759 ± 0.0000	0.6387 ± 0.0000
RF	**0.9654 ± 0.0005**	0.5231 ± 0.0106	0.9138 ± 0.0025	0.5761 ± 0.0100	0.7957 ± 0.0025	0.6626 ± 0.0049
XGB	0.9629 ± 0.0000	**0.5558 ± 0.0000**	**0.9169 ± 0.0000**	**0.6216 ± 0.0000**	**0.8142 ± 0.0000**	**0.7077 ± 0.0000**

In this experiment, the highest ROC AUC performance is achieved, however, domain shifts are not considered in the performance estimate. The threshold-dependent metrics for the RF for multiple thresholds are reported, which are determined by defining negative predictive values on the validation set (Table S3).

ii. assessed by random cross-validation with recent negatives.

The test set of experiment (ii) only comprises cases from the year 2020, which have been tested for COVID-19 with an RT-PCR test. Pre-pandemic negatives are excluded from the test set and the model is evaluated on pandemic samples only, which causes a performance drop from experiment (i) to (ii) (P = 0.016), see Table 2.

iii. assessed by temporal cross-validation.

In this experiment, the model is trained with samples until October and evaluated on samples from November and December. An additional performance drop in comparison to experiment (ii) (P = 0.016) is observed, which points to a domain shift over time which degrades predictive performance.

Mortality Prediction

iv. assessed by random cross-validation.

The samples are randomly shuffled and a five-fold nested cross-validation is performed. Again, the threshold-dependent metrics are reported (Table S6).

v. assessed by temporal cross-validation.

In this experiment, the model is trained with samples until October and evaluated on samples from November and December for mortality prediction of COVID-19 positive patients (*positives cohort*). The performance drops from experiment (iv) to (v) (P = 0.016), revealing a domain shift and over time for mortality prediction. The domain shifts over time again decrease the predictive performance.

The conducted experiments explore different levels of consideration of the domain shifts by different assessments. The evaluations are compared on the basis of ROC AUC as the PR AUC depends on the class prior, which varies in the different evaluation cohorts. The results expose the domain shifts and their diminishing effect on predictive

Table 3 Performance metrics of threshold-independent metrics for mortality prediction (experiment (iv)-(v)). The mean and the standard deviation (±) for the ROC AUC and PR AUC for the five random seeds are listed. Note that the PR AUC is dependent on the class prior, which changes with the different assessment strategies. The highest performance metrics per experiment are printed in bold

Model	**Experiment (iv)**		**Experiment (v)**	
	ROC AUC	PR AUC	ROC AUC	PR AUC
RE	0.5000 ± 0.0000	0.1592 ± 0.0351	0.5000 ± 0.0000	0.1320 ± 0.0000
BF	0.7599 ± 0.0748	0.4320 ± 0.1021	0.7483 ± 0.0000	0.3938 ± 0.0000
SNN	0.8656 ± 0.0356	0.5866 ± 0.1196	0.8478 ± 0.0053	0.4917 ± 0.0110
KNN	0.8207 ± 0.0550	0.5527 ± 0.1137	0.8272 ± 0.0000	0.4669 ± 0.0000
LR	0.8613 ± 0.0351	0.5555 ± 0.1281	0.8388 ± 0.0088	0.4784 ± 0.0173
SVM	0.8587 ± 0.0306	0.5679 ± 0.1010	0.8271 ± 0.0000	0.4185 ± 0.0001
RF	**0.8813 ± 0.0214**	**0.6267 ± 0.1065**	**0.8572 ± 0.0071**	**0.5556 ± 0.0127**
XGB	0.8501 ± 0.0210	0.5196 ± 0.1005	0.8038 ± 0.0000	0.4334 ± 0.0013

performance, as the performance drops from experiment (i) to (ii) and even further to (iii), and also from experiment (iv) to (v). By comparing experiment (i) and (ii) we investigate if inclusion of pre-pandemic negatives in the test set leads to overoptimistic metrics, and indeed variations in the performance metrics can be observed. We attribute this to the fact that the *2020 cohort* comprises patients who are suspect for COVID-19, some might even have characteristic symptoms, which are reflected in the blood tests. We hypothesize, that patients with characteristic symptoms tend to have similar blood test parameters, independent of their actual COVID-19 status. Therefore, a classification of the samples in the *2020 cohort* is more difficult and potential biases between the *2019* and *2020 cohort* cannot be exploited. Domain shifts over time within the year 2020 are considered in experiment (iii), which leads to a further decrease in predictive performance. Same holds for the drop of the predictive performance due to prospective evaluation in the mortality prediction task from experiment (iv) to (v).

Model Credibility under Domain Shifts

This experiment investigates the difference of the expected to the actual performance. The expected and actual results are compared for different simulated deployment times (June until December 2020) (Fig. 4). The expected performance is calculated on the respective preceding month (May until November). The expected ROC AUC is higher than the actual performance in most months (Fig. 4). The expected ROC AUC performance for December is significantly lower than the actual performance in December. The expected and actual PR AUC differ significantly in November and December. These results show the presence of a domain shift and thus there is a necessity for up-to-date assessments, otherwise the performance estimate is not trustworthy.

Credible and highly performant ML models for in-hospital applications require frequent re-training and re-assessments to combat the domain shift effects. Stronger weighting of more recent samples increases the predictive performance under domain shifts. More details on the methods and results to frequent re-training and stronger weighting of more recent samples are described in the Supplementary Information.

Discussion

Our set of experiments exposes the presence of domain shifts in COVID-19 blood test datasets as well as their detrimental effect on ML models. These domain shifts were insufficiently considered in previous works, which might have led to poor performance or even failure of the ML models in clinical practice. Therefore, our results suggest that the model performance should be frequently re-assessed. An up-to-date temporal evaluation appears indispensable to avoid unexpected behavior. The model should be frequently re-trained and more recent samples should be weighted stronger to exploit newly acquired samples and, thus, to counter the domain shift effect (see supplementary information, section Weighting of Recent Samples). Frequent re-training from scratch is a simple and feasible solution to handle the domain shifts, as ML models, such as RF or XGB for tabular data can easily be trained with limited computational resources. A high re-training frequency leads to fast adaptation to domain shifts and further to accurate predictions and assessments, but it is also associated with high effort for the acquisition of new samples and re-training of the ML models. This trade-off has to be balanced when selecting the re-training frequency in the hospital. Further, methods to handle the domain shifts could be considered, such as stronger weighting of recent samples during training.

In this large-scale study, we trained and evaluated our models with more samples than most studies [29–33] and we exploited pre-pandemic negative samples, which vastly increases our dataset size. The ML models achieved high predictive performance, comparable to previous studies [30–32, 35, 47], although the results cannot be directly compared as our assessment procedure is more rigorous. Different assessment procedures within our study also yielded highly variable performance estimates. In accordance with previous studies [29, 30, 35, 42, 48], XGB or RF for COVID-19 diagnosis and RF for mortality prediction were found to perform best. For increased validity and comparability of published performance estimates of clinical prediction models, it is highly recommended that authors stick to guidelines, such as TRIPOD-AI, thereby increasing the quality of published works in the medical AI research community.

One limitation of our work could be that we did not evaluate the generalization of our model to other hospitals. A transfer of a COVID-19 diagnostic model should only be done with thorough re-assessments, as a domain shift between hospitals might be present. However, this is not part of our investigation.

By automatic scanning of all blood tests, a large number of patients can be tested for COVID-19, which would not be feasible with expensive and slow RT-PCR tests. The ML predictions could enhance the established testing strategies in the hospitals, thereby broadening the screening. For re-training, at least some recent blood tests with associated ground truth RT-PCR test results have to be acquired to allow countering the domain shifts.

Our findings about domain shifts are not only relevant for COVID-19 datasets, but also transfer to other medical tasks, or in general, other applications of ML, where domain shifts occur. By advancing this field of research, we want to increase patient safety and protect clinical staff and we wish to make a contribution in banning the pandemic.

Abbreviations ACC: Accuracy; Adm. type: Type of hospital admission; BACC: Balanced accuracy; BF: Best feature; BMSGPK: Federal Ministry of Social Affairs, Health, Care and Consumer Protection; COVID-19: Coronavirus disease; c-statistic: Concordance statistic; CT: Computer tomography; F1: F1-score; i: Inpatient; KNN: K-nearest neighbor; LR: Logistic regression; MCH: Mean cell hemoglobin; MCHC: Mean corpuscular hemoglobin concentration; MCV: Mean corpuscular volume; ML: Machine learning; NPV: Negative predictive value; o: Outpatient; PPV: Positive predictive value; PR AUC: Precision recall area under the curve; RE: Random estimator; RF: Random forest; ROC AUC: Receiver operating characteristic area under the curve; RT-PCR: Reverse-transcriptome polymerase chain reaction; SNN: Self normalizing neural network; SVM: Support vector machine; XGB: Extreme gradient boosting

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10916-022-01807-1.

Acknowledgements We thank the projects Medical Cognitive Computing Center (Wi-2018-439843) and AI-MOTION (LIT-2018-6-YOU-212), AI-SNN (LIT-2018-6-YOU-214), Deep-Flood (LIT-2019-8-YOU-213), PRIMAL (FFG-873979), S3AI (FFG-872172), DL for granular flow (FFG-871302), ELISE (H2020-ICT-2019-3 ID: 951847), AIDD (MSCA-ITN-2020 ID: 956832). We thank Janssen Pharmaceutica, UCB Bio-pharma SRL, Merck Healthcare KGaA, Audi.JKU Deep Learning Center, TGW LOGISTICS GROUP GMBH, Silicon Austria Labs (SAL), FILL Gesellschaft mbH, Anyline GmbH, Google, ZF Friedrichshafen AG, Robert Bosch GmbH, Software Competence Center Hagenberg GmbH, TÜV Austria, and the NVIDIA Corporation. We thank Franz Grandits, Innosol for the daily download of the age distribution data of the newly infected COVID-19 patients from BMSGPK.

Authors' Contributions T.R., T.T., J.M., S.H. and G.K. designed the study. C.B. exported and anonymized the data from the hospital information system. T.R., A.M. and T.T. pre-processed the blood tests. T.R. pre-processed the RT-PCR tests and mortality data. T.R. implemented the ML models and conducted the experiments. T.R., S.H. and G.K. wrote the manuscript. T.T. wrote the application for the ethics approval. S.H., J.M. and G.K. supervised the project. All authors critically revised the draft and approved the final manuscript.

Funding Open access funding provided by Johannes Kepler University Linz. This work was supported by the projects Medical Cognitive Computing Center (Wi-2018–439843) and AI-MOTION (LIT- 2018–6-YOU-212).

Availability of Data and Material Data sharing requires ethics approval.

Code Availability Code to reproduce our study is provided on GitHub (https://github.com/ml-jku/covid).

Declarations

Ethics Approval Ethics approval for this study was obtained from the ethics committee of the Johannes Kepler University, Linz (approval number: 1104/2020). In our study, we analyze anonymized data only.

Conflicts of Interest The authors declare no conflicts of interest.

References

1. V. M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D. K. Chu, T. Bleicker, S. Brünink, J. Schneider, M. L. Schmidt, D. G. Mulders, B. L. Haagmans, B. Veer, S. Brink, L. Wijsman, G. Goderski, J. L. Romette, J. Ellis, M. Zambon, M. Peiris, H. Goossens, C. Reusken, M. P. G. Koopmans and C. Drosten, „Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR“, *Euro Surveill.,* 25, p. 2000045, 2020. https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045
2. N. Zhu, D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, P. Niu, F. Zhan, X. Ma, D. Wang, W. Xu, G. Wu, G. F. Gao and W. Tan, „A Novel Coronavirus from Patients with Pneumonia in China, 2019“, *N. Engl. J. Med.,* 382, pp. 727-733, 2020. https://doi.org/10.1056/NEJMoa2001017
3. G. C. Mak, P. K. Cheng, S. S. Lau, K. K. Wong, C. S. Lau, E. T. Lam, R. C. Chan and D. N. Tsang, „Evaluation of rapid antigen test for detection of SARS-CoV-2 virus“, *J. Clin. Virol.,* 129, p. 104500, 2020. https://doi.org/10.1016/j.jcv.2020.104500
4. M. J. Mina, R. Parker and D. B. Larremore, „Rethinking Covid-19 Test Sensitivity — A Strategy for Containment“, *N. Engl. J. Med.,* 383, p. e120, 2020. https://doi.org/10.1056/NEJMp2025631
5. E. T. Chin, B. Q. Huynh, L. A. C. Chapman, M. Murrill, S. Basu and N. C. Lo, „Frequency of Routine Testing for Coronavirus Disease 2019 (COVID-19) in High-risk Healthcare Environments to Reduce Outbreaks“, *Clin. Infect. Dis.,* p. ciaa1383, 2020. https://doi.org/10.1093/cid/ciaa1383
6. D. B. Larremore, B. Wilder, E. Lester, S. Shehata, J. M. Burke, J. A. Hay, M. Tambe, M. J. Mina and R. Parker, „Test sensitivity is secondary to frequency and turnaround time for COVID-19 surveillance“, *Sci. Adv.* 7, 2020. https://doi.org/10.1126/sciadv.abd5393
7. M. van der Schaar, A. M. Alaa, A. Floto, A. Gimson, S. Scholtes, A. Wood, E. McKinney, D. Jarrett, P. Lio A. Ercole, “How artificial intelligence and machine learning can help healthcare systems respond to COVID-19”, *Mach. Learn.,* 110, 1–14, 2021. https://doi.org/10.1007/s10994-020-05928-x
8. T. Alafif, A. M. Tehame, S. Bajaba, A. Barnawi and S. Zia, „Machine and Deep Learning towards COVID-19 Diagnosis and Treatment: Survey, Challenges, and Future Directions“, *Int. J. Environ. Res. Public Health,* 18, pp. 1-24, 2021. https://doi.org/10.3390/ijerph18031117
9. A. S. Adly, A. S. Adly and M. S. Adly, „Approaches Based on Artificial Intelligence and the Internet of Intelligent Things to Prevent the Spread of COVID-19: Scoping Review“, *J. Med. Internet Res.,* 22, 8, p. e19104, 8 2020. https://doi.org/10.2196/19104
10. A. Majeed, and S. Lee, “Applications of Machine Learning and High- Performance Computing in the Era of COVID-19”, *Appl. Syst. Innov.*, 4, 40, 2021, https://doi.org/10.3390/asi4030040

11. A. Bansal, R.P. Padappayil, C. Garg, A. Singal, M. Gupta and A. Klein, “Utility of Artificial Intelligence Amidst the COVID 19 Pandemic: A Review”, *J. Med. Syst.*, 44**,** 156, 2020. https://doi.org/10.1007/s10916-020-01617-3
12. C.M. Williams, R. Chaturvedi, R.D. Urman, R.S. Waterman and R.A. Gabriel, “Artificial Intelligence and a Pandemic: an Analysis of the Potential Uses and Drawbacks”, *J. Med. Syst.,* 45**,** 26, 2021. https://doi.org/10.1007/s10916-021-01705-y
13. A.S. Albahri, R.A. Hamid, J.k. Alwan, Z.T. Al-qays, A.A. Zaidan, B.B. Zaidan, A.O.S. Albahri, A.H. AlAmoodi, J.M. Khlaf, E.M. Almahdi, E. Thabet, S.M. Hadi, K.I. Mohammed, M.A. Alsalem, J.R. Al-Obaidi and H.T. Madhloom, “Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review”, *J. Med. Syst.,* 44**,** 122, 2020. https://doi.org/10.1007/s10916-020-01582-x
14. A. K. Arshadi, J. Webb, M. Salem, E. Cruz, S. Calad-Thomson, N. Ghadirian, J. Collins, E. Diez-Cecilia, B. Kelly, H. Goodarzi and J. S. Yuan, „Artificial Intelligence for COVID-19 Drug Discovery and Vaccine Development“, *Front. Artif. Intell. Appl.,* 3, p. 65, 2020. https://doi.org/10.3389/frai.2020.00065
15. E. Ong, M. U. Wong, A. Huffman and Y. He, „COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning“, *Front. Immunol.,* 11, p. 1581, 2020. https://doi.org/10.3389/fimmu.2020.01581
16. M. Hofmarcher, A. Mayr, E. Rumetshofer, P. Ruch, P. Renz, J. Schimunek, P. Seidl, A. Vall, M. Widrich, S. Hochreiter and G. Klambauer, „Large-scale ligand-based virtual screening for SARS-CoV-2 inhibitors using deep neural networks“, *arXiv,* pp. 2010.06498v2, preprint: not peer reviewed, 2021. arXiv: 2004.00979
17. I. Ozsahin, B. Sekeroglu, M. S. Musa, M. T. Mustapha and D. U. Ozsahin, „Review on Diagnosis of COVID-19 from Chest CT Images Using Artificial Intelligence“, *Comput. Math. Method. M.,* 2020, p. 1-10, 2020. https://doi.org/10.1155/2020/9756518
18. A. A. Borkowski, N. A. Viswanadhan, L. B. Thomas, R. D. Guzman, L. A. Deland and S. M. Mastorides, „Using Artificial Intelligence for COVID-19 Chest X-ray Diagnosis“, *Fed. Pract.,* 37, p. 398–404, 2020. https://doi.org/10.12788/fp.0045
19. P. Saha, M. S. Sadi and M. M. Islam, „EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers“, *Inform. Med. Unlocked,* 22, p. 100505, 2021. https://doi.org/10.1016/j.imu.2020.100505
20. T. D. Pham, „Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning?“, *Health inf. sci. syst.,* 9, p. 1-11, 2021. https://doi.org/10.1007/s13755-020-00135-3
21. S. Tabik, A. Gómez-Ríos, J. L. Martín-Rodríguez, I. Sevillano-García, M. Rey-Area, D. Charte, E. Guirado, J. L. Suárez, J. Luengo, M. A. Valero-González, P. García-Villanova, E. Olmedo-Sánchez and F. Herrera, „COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images“, *IEEE J. Biomed. and Health Inform.,* 24, p. 3595-3605, 2020. https://doi.org/10.1109/JBHI.2020.3037127
22. G. Wang, X. Liu, J. Shen, C. Wang, Z. Li, L. Ye, X. Wu, T. Chen, K. Wang, X. Zhang, Z. Zhou, J. Yang, Y. Sang, R. Deng, W. Liang, T. Yu, M. Gao, J. Wang, Z. Yang, H. Cai, G. Lu, L. Zhang, L. Yang, W. Xu, W. Wang, A. Olevera, I. Ziyar, C. Zhang, O. Li, W. Liao, J. Liu, W. Chen, W. Chen, J. Shi, L. Zheng, L. Zhang, Z. Yan, X. Zou, G. Lin, G. Cao, L. L. Lau, L. Mo, Y. Liang, M. Roberts, E. Sala, C.-B. Schönlieb, M. Fok, J. Yiu-Nam Lau, T. Xu, J. He, K. Zhang, W. Li, T. Lin, “A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images”, *Nat. Biomed. Eng.*, 5, p. 509–521, 2021. https://doi.org/10.1038/s41551-021-00704-1
23. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, AIX-COVNET, J. H. F. Rudd, E. Sala & C.-B. Schönlieb, “Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans”, *Nat. Mach. Intell.*, 2021, https://doi.org/10.1038/s42256-021-00307-0
24. K. Hammoudi, H. Benhabiles, M. Melkemi, F. Dornaika, I. Arganda-Carreras, D. Collard and A. Scherpereel, “Deep Learning on Chest X-ray Images to Detect and Evaluate Pneumonia Cases at the Era of COVID-19”, *J. Med. Syst.*, 45**,** 75, 2021. https://doi.org/10.1007/s10916-021-01745-4
25. K. Santosh, S. Ghosh, “Covid-19 Imaging Tools: How Big Data is Big?”, *J. Med. Syst.*, 45**,** 71, 2021. https://doi.org/10.1007/s10916-021-01747-2
26. P. Mouawad, T. Dubnov and S. Dubnov, „Robust Detection of COVID-19 in Cough Sounds“, *SN Computer Science*, 2, p. 34, 2021. https://doi.org/10.1007/s42979-020-00422-6
27. B. W. Schuller, D. M. Schuller, K. Qian, J. Liu, H. Zheng and X. Li, „COVID-19 and Computer Audition: An Overview on What-Speech & Sound Analysis Could Contribute in the SARS-CoV-2 Corona Crisis“, *Front. Digit. Health,* 2020. https://doi.org/10.3389/fdgth.2021.564906
28. J. Laguarta, F. Hueto and B. Subirana, „COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings“, *IEEE open j. eng. med. biol.*, 1, p. 275-281, 2020. https://doi.org/10.1109/OJEMB.2020.3026928
29. T. Tschoellitsch, M. Dünser, C. Böck, K. Schwarzbauer and J. Meier, „Machine Learning Prediction of SARS-CoV-2 Polymerase Chain Reaction Results with Routine Blood Tests“, *Lab. Med.,* 52, p. 146-149, 2020. https://doi.org/10.1093/labmed/lmaa111
30. D. Brinati, A. Campagner, D. Ferrari, M. Locatelli, G. Banfi and F. Cabitza, „Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study“, *J. Med. Syst.,* 44, p. 135, 2020. https://doi.org/10.1007/s10916-020-01597-4
31. D. Goodman-Meza, A. Rudas, J. N.Chiang, P. C. Adamson, J. Ebinger, N. Sun, P. Botting, J. A. Fulcher, F. G. Saab, R. Brook, E. Eskin, U. An, M. Kordi, B. Jew, B. Balliu, Z. Chen, B. L. Hill, E. Rahman, E. Halperin and V. Manue, „A machine learning algorithm to increase COVID-19 inpatient diagnostic capacity“, *Plos One,* 15, p. e0239474, 2020. https://doi.org/10.1371/journal.pone.0239474
32. F. Cabitza, A. Campagner, D. Ferrari, C. D. Resta, D. Ceriotti, E. Sabetta, A. Colombini, E. D. Vecchi, G. Banfi, M. Locatelli and A. Carobene, „Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests“, *Clin. Chem. Lab. Med.,* 59, p. 421-431, 2021. https://doi.org/10.1515/cclm-2020-1294
33. T. Langer, M. Favarato, R. Giudici, G. Bassi, R. Garberi, F. Villa, H. Gay, A. Zeduri, S. Bragagnolo, A. Molteni, M. C. Andrea Beretta, M. Moreno, C. Vismara, C. F. Perno, M. Buscema, E. Grossi and R. Fumagalli, „Development of machine learning models to predict RT-PCR results for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in patients with influenza-like symptoms using only basic clinical data“, *Scand. j. trauma resusc.,* 28, p. 1–14, 2020. https://doi.org/10.1186/s13049-020-00808-8
34. A. A. S. Soltan, S. Kouchaki, T. Zhu, D. Kiyasseh, T. Taylor, Z. B. Hussain, T. Peto, A. J. Brent, D. W. Eyre and D. A. Clifton, „Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test“, *Lancet Digit. Health,* 3, p. 78-87, 2021. https://doi.org/10.1016/s2589-7500(20)30274-0
35. H. S. Yang, Y. Hou, L. V. Vasovic, P. A. D. Steel, A. Chadburn, S. E. Racine-Brzostek, P. Velu, M. M. Cushing, M. Loda, R. Kaushal, Z. Zhao and F. Wang, „Routine Laboratory Blood Tests Predict SARS-CoV-2 Infection Using Machine Learning“, *Clin. Chem.,* 66, p. 1396-1404, 2020. https://doi.org/10.1093/clinchem/hvaa200

36. T. B. Plante, A. M. Blau, A. N. Berg, A. S. Weinberg, I. C. Jun, V. F. Tapson, T. S. Kanigan and A. B. Adib, „Development and External Validation of a Machine Learning Tool to Rule Out COVID-19 Among Adults in the Emergency Department Using Routine Blood Tests: A Large, Multicenter, Real-World Study“, *J. Med. Internet Res.,* 22, p. 1-19, 2020. https://doi.org/10.2196/24048

37. M. Almansoor and N. M. Hewahi, „Exploring the Relation between Blood Tests and Covid-19 Using Machine Learning“, *ICDABI,* p. 1–6, 2020. https://doi.org/10.1109/ICDABI51230.2020.9325673

38. M. AlJame, I. Ahmad, A. Imtiaz and A. Mohammed, „Ensemble learning model for diagnosing COVID-19 from routine blood tests“, *Inform. Med. Unlocked,* 21, p. 100449, 2020. https://doi.org/10.1016/j.imu.2020.100449

39. V. Formica, M. Minieri, S. Bernardini, M. Ciotti, C. D'Agostini, M. Roselli, M. Andreoni, C. Morelli, G. Parisi, M. Federici, C. Paganelli and J. M. Legramante, „Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2“, *Clin. Med.,* 20, p. e114--e119, 2020. https://doi.org/10.7861/clinmed.2020-0373

40. V. A. Freitas Barbosa, J. C. Gomes, M. A. Santana, J. E. A. Albuquerque, R. G. Souza, R. E. Souza and W. P. Santos, „Heg. IA: an intelligent system to support diagnosis of Covid-19 based on blood tests“, *Res. Biomed. Eng.,* 2021. https://doi.org/10.1007/s42600-020-00112-5

41. A. Banerjee, S. Ray, B. Vorselaars, J. Kitson, M. Mamalakis, S. Weeks, M. Baker and L. S. Mackenzie, „Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population“, *Int. Immunopharmacol.,* 86, 2020. https://doi.org/10.1016/j.intimp.2020.106705

42. E. C. Silveira, „Prediction of COVID-19 From Hemogram Results and Age Using Machine Learning“, *Front. health inform.,* 9, p. 39, 2020. https://doi.org/10.30699/fhi.v9i1.234

43. E. Avila, A. Kahmann, C. Alho and M. Dorn, „Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios“, *PeerJ,* 8, p. e9482, 2020. https://doi.org/10.7717/peerj.9482

44. H. Sun, A. Jain, M. J. Leone, H. S. Alabsi, L. N. Brenner, E. Ye, W. Ge, Y.-P. Shao, C. L. Boutros, R. Wang, R. A. Tesh, C. Magdamo, S. I. Collens, W. Ganglberger, I. V. Bassett, J. B. Meigs, J. Kalpathy-Cramer, M. D. Li, J. T. Chu, M. L. Dougan, L. W. Stratton, J. Rosand, B. Fischl, S. Das, S. S. Mukerji, G. K. Robbins and M. B. Westover, „CoVA: An Acuity Score for Outpatient Screening that Predicts Coronavirus Disease 2019 Prognosis“, *J. Infect. Dis.,* 223, p. 38-46, 2020. https://doi.org/10.1093/infdis/jiaa663

45. Y. Zhou, B. Li, J. Liu and D. Chen, „The Predictive Effectiveness of Blood Biochemical Indexes for the Severity of COVID-19“, *Can. J. Infect. Dis. Med. Microbiol.,* 2020, p. 732081, 2020. https://doi.org/10.1038/s41598-021-83784-y

46. A. L. Booth, E. Abels and P. McCaffrey, „Development of a prognostic model for mortality in COVID-19 infection using machine learning“, *Mod. Pathol.,* 34, p. 522-531, 2020. https://doi.org/10.1038/s41379-020-00700-x

47. X. Li, P. Ge, J. Zhu, H. Li, J. Graham, A. Singer, P. S. Richman and T. Q. Duong, „Deep learning prediction of likelihood of ICU admission and mortality in COVID-19 patients using clinical variables“, *PeerJ,* 8, p. e10337, 2020. https://doi.org/10.7717/peerj.10337

48. F. S. Heldt, M. P. Vizcaychipi, S. Peacock, M. Cinelli, L. McLachlan, F. Andreotti, S. Jovanović, N. L. Robert Dürichen, R. A. Fletcher, A. Hancock, A. McCarthy, R. A. Pointon, A. Brown, J. Eaton, R. Liddi, L. Mackillop, L. Tarassenko and R. T. Khan, „Early risk assessment for COVID-19 patients from emergency department data using machine learning“, *Sci. Rep.,* 11, p. 4200, 2021. https://doi.org/10.1038/s41598-021-83784-y

49. S. Heber, D. Pereyra, W. C. Schrottmaier, K. Kammerer, J. Santol, E. Pawelka, M. Hana, A. Scholz, M. Liu, A. Hell, K. Heiplik, B. Lickefett, S. Havervall, M. T. Traugott, M. Neuböck, C. Schörgenhofer, T. Seitz, C. Firbas, M. Karolyi, G. Weiss, B. Jilma, C. Thralin, R. Bellmann-Weiler, H. J. F. Salzer, M. J. M. Fischer, A. Zoufaly and A. Assinger, "Development and external validation of a logistic regression derived formula based on repeated routine hematological measurements predicting survival of hospitalized Covid-19 patients“, *medRxiv,* 2020. https://doi.org/10.1101/2020.12.20.20248563

50. Y. Gao, G.-Y. Cai, W. Fang, H.-Y. Li, S.-Y. Wang, L. Chen, Y. Yu, D. Liu, S. Xu, P.-F. Cui, S.-Q. Zeng, X.-X. Feng, R.-D. Yu, Y. Wang, Y. Yuan, X.-F. Jiao, J.-H. Chi, J.-H. Liu, R.-Y. Li, X. Zheng, C.-Y. Song, N. Jin, W.-J. Gong, X.-Y. Liu, L. Huang, X. Tian, L. Li, H. Xing, D. Ma, C.-R. Li, F. Ye and Q.-L. Gao, "Machine learning based early warning system enables accurate mortality risk prediction for COVID-19“, *Nat. Commun.,* 11, p. 5033, 2020. https://doi.org/10.1038/s41467-020-18684-2

51. A. Vaid, S. Somani, A. J. Russak, J. K. De Freitas, F. F. Chaudhry, I. Paranjpe, K. W. Johnson, S. J. Lee, R. Miotto, F. Richter, S. Zhao, N. D. Beckmann, N. Naik, A. Kia, P. Timsina, A. Lala, M. Paranjpe, E. Golden, M. Danieletto, M. Singh, D. Meyer, P. F. O\'Reilly, L. Huckins, P. Kovatch, J. Finkelstein, R. M. Freeman, E. Argulian, A. Kasarskis, B. Percha, J. A. Aberg, E. Bagiella, C. R. Horowitz, B. Murphy, E. J. Nestler, E. E. Schadt, J. H. Cho, C. Cordon-Cardo, V. Fuster, D. S. Charney, D. L. Reich, E. P. Bottinger, M. A. Levin, J. Narula, Z. A. Fayad, A. C. Just, A. W. Charney, G. N. Nadkarni and B. S. Glicksberg, „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, p. 1–19, 2020. https://doi.org/10.2196/24018

52. H. Ko, H. Chung, W. S. Kang, C. Park, D. W. Kim, S. E. Kim, C. R. Chung, R. E. Ko, H. Lee, J. H. Seo, T.-Y. Choi, R. Jaimes, K. W. Kim and J. Lee, „An Artificial Intelligence Model to Predict the Mortality of COVID-19 Patients at Hospital Admission Time Using Routine Blood Samples: Development and Validation of an Ensemble Model“, *J. Med. Internet Res.,* 22, p. e25442, 2020. https://doi.org/10.2196/25442

53. K. Stacke, G. Eilertsen, J. Unger, C. Lundstrom, "Measuring Domain Shift for Deep Learning in Histopathology". *IEEE J. Biomed. Health Inform.*, 25(2), p.325-336, 2021 https://doi.org/10.1109/JBHI.2020.3032060

54. P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn and P. Liang, „WILDS: A Benchmark of in-the-Wild Distribution Shifts“, *Proceedings of Machine Learning Research,* 139:5637–5664, 2021. arXiv: 2012.07421

55. J. J. Thiagarajan, R. Deepta, P. Sattigeri, "Understanding Behavior of Clinical Models under Domain Shifts", *arXiv,* p. 1809.07806v2, 2019, preprint: not peer reviewed., arXiv: 1809.07806

56. M. Schneider, L. Wang, C. Marr, "Evaluation of Domain Adaptation Approaches for Robust Classification of Heterogeneous Biological Data Sets", *Artificial Neural Networks and Machine Learning -- ICANN 2019: Deep Learning*, pp. 673—686, 2019, ISBN: 978–3–030–30484–3

57. G.S. Collins, P. Dhiman, C. L. Andaur Navarro, J. Ma , L. Hooft, J. B. Reitsma, P. Logullo, A. L. Beam , L. Peng, B. Van Calster , M. van Smeden, R. D. Riley, K. G. M. Moons, "Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence", *BMJ Open* 11:e048008. 2021 https://doi.org/10.1136/bmjopen-2020-048008

58. Federal Ministry of Social Affairs, Health, Care and Consumer Protection (BMSGPK), *Austrian COVID-19 Open Data Information portal.* https://www.data.gv.at/covid-19/

59. M. Wulfmeier, A. Bewley and I. Posner, „Incremental Adversarial Domain Adaptation for Continually Changing Environments“, *ICRA,* p. 1–9, 2018. https://doi.org/10.1109/ICRA.2018.8460982
60. T. Adler, J. Brandstetter, M. Widrich, A. Mayr, D. Kreil, M. Kopp, G. Klambauer and S. Hochreiter, „Cross-Domain Few-Shot Learning by Representation Fusion,“ *arXiv,* p. 2010.06498v2, preprint: not peer reviewed., 2021. arXiv: 2010.06498
61. V. L. S. Crawford, O. Sweeney, P. V. Coyle, I. M. Halliday and R. W. Stout, „The relationship between elevated fibrinogen and markers of infection: a comparison of seasonal cycles“, *QJM: An International Journal of Medicine,* 93, p. 745–750, 2000. https://doi.org/10.1093/qjmed/93.11.745
62. H. Elsahar and M. Gallé, „To Annotate or Not? Predicting Performance Drop under Domain Shift“, *EMNLP-IJCNLP,* 9, p. 2163-2173, 2019. https://doi.org/10.18653/v1/D19-1222
63. G. Klambauer, T. Unterthiner, A. Mayr and S. Hochreiter, „Self-normalizing neural networks“, *NIPS,* p. 971–980, 2017. arXiv:1706.02515
64. F. E. Harrell Jr., K. L. Lee and D. B. Mark, „Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors“, *Stat. Med.,* 15, p. 361-387, 1996. https://doi.org/10.1002/0470023678.CH2B(I)
65. J. Davis and M. Goadrich, „The relationship between Precision-Recall and ROC Curves“, *ICML,* 23, p. 233–240, 2006. https://doi.org/10.1145/1143844.1143874
66. P. Branco, L. Torgo and R. P. Ribeiro, „A Survey of Predictive Modeling on Imbalanced Domains“, *ACM Comput. Surv.,* 49, p. 1-50, 2016. https://doi.org/10.1145/2907070
67. J. H. Zar, “Biostatistical Analysis”, 5th edition, Prentice Hall, Upper Saddle River, New Jersey USA, 2010, ISBN: 9780321656865