European Journal of Radiology Open 11 (2023) 100497

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Performance of an AI algorithm during the different phases of the COVID pandemics: what can we learn from the AI and vice versa.

Michele Catalano [a], Chandra Bortolotto [a], Giovanna Nicora [b], Marina Francesca Achilli [a,*,1], Alessio Consonni [a], Lidia Ruongo [a], Giovanni Callea [a], Antonio Lo Tito [a], Carla Biasibetti [a], Antonella Donatelli [a], Sara Cutti [c], Federico Comotto [d], Giulia Maria Stella [h], Angelo Corsico [h], Stefano Perlini [e], Riccardo Bellazzi [b], Raffaele Bruno [f], Andrea Filippi [g], Lorenzo Preda [a]

[a] *Department of Clinical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy and Radiology Department, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy*
[b] *Department of Electrical, Computer and Biomedical Engineering, University of Pavia, Pavia, Italy*
[c] *Medical Direction, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy*
[d] *Reply S.p.A., Corso Francia, 110, Turin, Italy*
[e] *Department of Internal Medicine and Therapeutics, University of Pavia, Pavia, Italy and Dept. of Emergency Department, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy*
[f] *Department of Clinical, Surgical, Diagnostic and Pediatric Sciences, University of Pavia, Pavia, Italy and Infectious Diseases Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy*
[g] *Radiation Oncology Unit, University of Pavia, Pavia, Italy and Infectious Diseases Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy*
[h] *Department of Internal Medicine and Therapeutics, University of Pavia, Pavia, Italy and Dept. of Respiratory Diseases Unit, Fondazione IRCCS Policlinico San Matteo, Pavia, Italy*

ARTICLE INFO

Keywords:
Artificial intelligence
COVID 19
X-rays
Machine learning
Resource allocation

ABSTRACT

Background: Artificial intelligence (AI) has proved to be of great value in diagnosing and managing *Sars-Cov-2* infection. ALFABETO (ALL-FAster-BEtter-TOgether) is a tool created to support healthcare professionals in the triage, mainly in optimizing hospital admissions.
Methods: The AI was trained during the pandemic's "first wave" (February-April 2020). Our aim was to assess the performance during the "third wave" of the pandemics (February-April 2021) and evaluate its evolution. The neural network proposed behavior (hospitalization vs home care) was compared with what was actually done. If there were discrepancies between ALFABETO's predictions and clinicians' decisions, the disease's progression was monitored. Clinical course was defined as "favorable/mild" if patients could be managed at home or in spoke centers and "unfavorable/severe" if patients need to be managed in a hub center.
Results: ALFABETO showed accuracy of 76%, AUROC of 83%; specificity was 78% and recall 74%. ALFABETO also showed high precision (88%). 81 hospitalized patients were incorrectly predicted to be in "home care" class. Among those "home-cared" by the AI and "hospitalized" by the clinicians, 3 out of 4 misclassified patients (76.5%) showed a favorable/mild clinical course. ALFABETO's performance matched the reports in literature.
Conclusions: The discrepancies mostly occurred when the AI predicted patients could stay at home but clinicians hospitalized them; these cases could be handled in spoke centers rather than hubs, and the discrepancies may aid clinicians in patient selection. The interaction between AI and human experience has the potential to improve both AI performance and our comprehension of pandemic management.

Abbreviations: ALFABETO, (ALL-FAster-BEtter-TOgether); AI, artificial intelligence; ML, Machine learning; DL, deep learning; ER, emergency room; RT-PCR, real time polymerase chain reaction; COPD, chronic obstructive pulmonary disease; WBC, white blood count; CXR, chest X ray; CRP, C-reactive Protein; SaO2, oxygen saturation; AUROC, Area Under the Curve of Receiver Characteristic Operator; RF, Random Forest.

* Corresponding author.
E-mail address: marinafrancesc.achilli01@universitadipavia.it (M.F. Achilli).
[1] ORCID: https://orcid.org/0009-0008-4892-468X

https://doi.org/10.1016/j.ejro.2023.100497
Received 6 May 2023; Received in revised form 2 June 2023; Accepted 4 June 2023
Available online 19 June 2023

1. Background

Since its identification in December 2019 in Wuhan, China, and subsequent declaration as pandemic on 11 March 2020, COVID-19 has been putting an unprecedented and increasing strain on healthcare resources, hard testing our ability to deliver effective healthcare, especially during the first wave of pandemics (February – April 2020) [1,2].

Artificial intelligence (AI) has raised hope to be of great value in diagnosing and managing COVID infection [3,4,5,6]. AI is a scientific area whose aim is simulating human intellectual processes through technological devices. Machine learning (ML) is a subcategory of AI that creates systems that can learn tasks or predict future outcomes relying on pre-processed data. Among the different ML methods, deep learning (DL) has the capability to learn and represent features automatically; this eliminates the need to manually engineer features based on human expertise and hence obtain higher accuracies for different classification and regression tasks. Although ML and DL-based methods have been successful in solving various problems, yet they suffer from two main problems. First, they require a large amount of training and can be computationally demanding. Moreover, they may struggle with generalizing effectively across different populations, which can lead to reduced performance in medicine when applied to diverse patient sets [7].

Since the pandemic began in early 2020 numerous AI systems for diagnosis and prognosis of COVID-19 using radiological imaging have been developed and hundreds of manuscripts have been written [6].

While this huge potential, the successful practical deployments of these AI-based tools have been limited by several challenges such as limited data accessibility, need for external evaluation of AI models, lack of awareness of AI experts of the regulatory landscape governing the deployment of AI tools in healthcare, the need for clinicians and other experts to work with AI experts in a multidisciplinary context and the need to address public concerns over data collection, privacy and protection [8]. With no standardization, AI algorithms for COVID-19 have been developed from a very broad range of applications, data collection procedures and performance assessment metrics. As a result, none are currently ready to be deployed clinically [6]. One of the key challenges that AI experts have faced during the COVID-19 pandemic is the lack of access to sufficiently large datasets for training and external validation of AI models upon which deployable and successful applications depend [9]. In some cases, this led to work on so-called "Frankenstein datasets" with algorithms being trained and tested on identical or overlapping datasets while believing them to be from distinct sources [6]. Moreover, many of the developed AI-based techniques and models for COVID-19 diagnosis and epidemiological forecasting have not been externally evaluated. External model evaluation helps in assessing the generalizability of the predictions on independent datasets and ensures that the model has learnt the underlying features of the process that produces the data rather than "memorized" the features of a particular set of data [8].

The COVID-19 pandemic has highlighted the importance of domain specific knowledge in AI. It is important for the clinicians to work with AI experts to help them understand the context of the solutions being developed, to help them interpret the results and to guide them on how those solutions could be used and integrated into existing clinical healthcare pathways or workflows [8].

Moreover, the absence and lack of engagement of clinicians to contribute and review research results during the COVID-19 pandemic has contributed to the limited impact, reliability and clinical utility of many of these research findings.

ALL FAster BEtter Together (ALFABETO) is an AI tool created to support healthcare professionals in the triage phase through predictive analyses on the evolution of the pathology, thus supporting the practitioners in the resource allocation. ALFABETO is based on ML technology and integrates heterogeneous and labeled data sources such as clinical, laboratoristic and imaging findings to develop predictive analyses of the possible short-medium term evolution of the pathology, providing the essential elements to decide the best assistance strategy (home care vs hospitalization). The project, publicly financed, was born from a joint venture between University, Hospitals and private companies specialized in the development of Artificial Intelligence solutions. ALFABETO aim is to optimize hospital admissions by suggesting, in case of uncertainty at triage, whether the Patient actually needs hospitalization over home care. The software was initially supervised-trained during the first wave of the pandemic (February-April 2020). The AI model predicts hospital admissions based on both clinical features and DL-extracted features from chest radiograph images.

Since literature stressed the need of performance monitoring and optimization for AI-driven tools able to generalize over time, our aim was to re-assess ALFABETO performance during the third wave of the pandemic (February-April 2021).

2. Methods

2.1. Participants

Data from 462 Patients of the third wave were used to re-assess the performance of ALFABETO. Clinical (e.g. dyspnea), anamnestic (such as comorbidities e.g. presence of cardiac disease, COPD, stroke), laboratory (e.g. WBC, CRP) data along with chest X ray performed at the emergency room's admission for each patient.

The inclusion criteria for the patients of this retrospective study were the following: patients over 16 years old; admission to the emergency room (ER) between March and April 2021 of a COVID-hub hospital; *Sars-CoV-2* infection confirmed by real time RT-PCR on nasopharyngeal swab; patients who underwent CXR examination. The negative result of the RT-PCR was an exclusion criterion for patients admitted in the emergency department. The absence of fever or respiratory symptoms was not considered an exclusion criterion.

The project received institutional review board approval and informed consent was obtained accordingly.

2.2. Source of data

2.2.1. Image acquisition (Chest X Ray, CXR)

We collected the CXR image performed at the emergency room's admission for each patient. All images were obtained through digital, baseline, frontal (either antero-posterior or postero-anterior) acquisitions.

2.2.2. Clinical/laboratory data, comorbidities and outcome

We collected all the clinical and laboratory data of the admission day for each patient. The clinical data were sex of the patient, age, body basal temperature (BBT), ambient air oxygen saturation levels (SaO2), presence of respiratory symptoms (cough and dyspnea), number of days with symptoms before the ER admission (giving a score of 1 from 0 to 2 days, of 2 from 3 to 5 days, of 3 from 6 to 9 days and 4 for more than 9 days) using the same timing subgrouping as Bernhaim and colleagues [10]. The laboratory data were White Blood Cells count (WBC) and C-reactive Protein (CRP) blood levels. All these parameters were selected by a Multidisciplinary Task Force as the most suitable clinical and laboratory data for assessing the severity of the disease combining a review of the available literature and clinical experience. We selected the following comorbidities from the past medical history of the patients: hypertension; type 2 diabetes mellitus; active cancer in the last 5 years; dementia; chronic obstructive pulmonary disease; chronic respiratory failure; chronic renal failure; obesity. We selected these comorbidities as they are associated with an increased risk of death in Sars-Cov2 patients in a large epidemiological survey (Epicentro, website at: https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza-dati). Since ALFABETO goal is to decide the best assistance strategy (home care vs hospitalization) and to optimize hospital admissions we decided to classify the patients' outcome in three categories: mild (not

hospitalized or hospitalized without need for ventilatory support); moderate (hospitalized with continuous positive airway pressure device support - CPAP); severe (hospitalized with invasive ventilatory support or deceased). The source of the raw data was the PACS (Picture archiving and communication system) of the Hospital in which the CXR images were stored; for clinical and laboratory information the source of the data was the HIS (Hospital information System).

2.2.3. AI structure

Using data collected during the first wave of the pandemic, we trained different supervised ML classifiers (Logistic Regression, Multilayer Perceptron, Gradient Boosting, Bayesian Network and Random Forest) to predict whether each patient experiencing triage should be hospitalized or not. ML classifiers for COVID-19 hospital admission are trained on clinical and laboratory features, as well as on features extracted from CXR images through DL. The DL approach used, named X-RAIS, was developed by one of the ALFABETO consortium partners.

X-RAIS is a deep network able to analyze different types of medical images and to extract relevant information for diagnosis. X-RAIS integrates a Deep Learning module dedicated to the analysis of radiological images and used by ALFABETO for the inspection of chest X-ray examinations. The input of this module is an image extracted directly from the DICOM file to which an appropriate pre-processing pipeline is applied. By analyzing the image, X-RAIS predicts the score associated with the presence of 5 different radiological signs; this score represents how high or low the radiological sign is present in the analyzed chest x-rays [0 =absence, 1 = maximum presence]. The 5 radiological signs are: Consolidation, Infiltration, Edema, Effusion, and Lung Opacity. The Deep Learning model of X-RAIS is based on a DenseNet121 trained and fine-tuned on biomedical images annotated with the 5 different radiological signs. Through its neural network layers, the model is able to recognize and extract hidden patterns from images and convert them into numerical values (vectors referring to the presence of the aforementioned radiological signs). Mimicking the expert eye of a radiologist, X-RAIS is able to identify these 5 categories and predict the related level of presence/absence. The algorithm was implemented using PyTorch open-source framework.

In this context, X-RAIS transforms the CRX image of a patient into 5 numerical relevant features: Consolidation, Infiltration, Edema, Effusion and Lung Opacity. To build the feature set for each patient, the 5 numerical features extracted from the patient's image were stacked with the features collected by clinicians and described in paragraph 2.2.2. The final features set is composed of 32 features, both numerical and categorical. Fig. 1 shows ALFABETO workflow for classification: clinical features and DL-extracted features are combined and provided as input to the classifier, whose output for each patient is the predicted class ("Home" vs "Hospital"), as well as the probability of classification. Using as reference training data, we remove features with more than 90% of missing values. We then impute missing data by using the most frequent value for categorical variables and the average for numerical variables. Data were then scaled between 0 and 1. To maximize recall (i.e., the ability to correctly identify hospitalized patients) and precision (i.e., the fraction patients needed hospitalization among all the patients predicted in the "Hospital" class), we selected the best threshold for classification based on cross validation results on training data to maximize the F score measure, which is the harmonic mean between precision and recall. The classification pipeline was implemented in Python 3.7 and code is available on Github (https://github.com/GiovannaNicora/ALFABETO). Among the different ML classifiers tested, the Random Forest (RF) showed better performance in terms of precision and recall. Therefore, we focus our analysis on predictions made by RF. RFs are ensemble classifiers based on multiple decision trees widely used in clinical application given their high performances in different prediction tasks [11,12].

All methods were carried out in accordance with relevant guidelines and regulations.

3. Results

3.1. Patient characteristics

A total of 462 patients tested positive for *Sars-CoV-2* by real time RT-PCR assay were considered in our study; among them, 196 (42.4%) were women and 266 (57.6%) were men. The average age was 67.7 ± 15.3

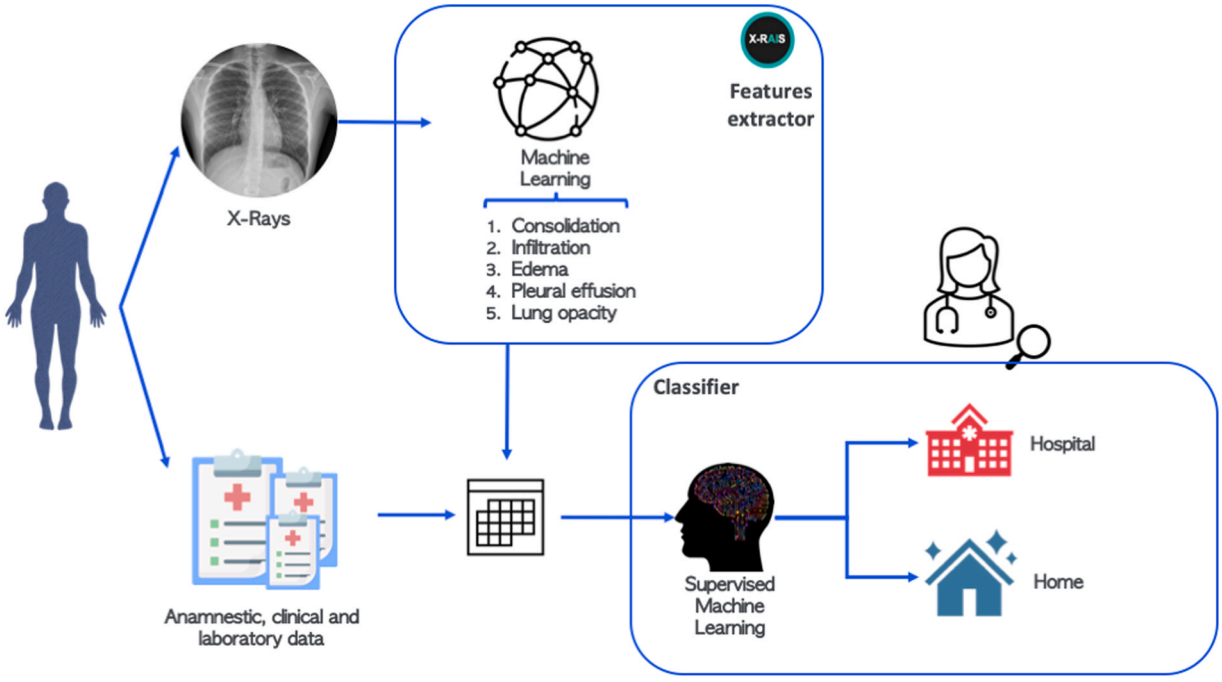

Fig. 1. A Conceptual view of ALFABETO Model. Clinical data and deep learning features extracted from CXR are provided as input for a machine learning classifier, which then predicts whether a specific patient should be hospitalized or not.

years (range 17–100 years). When available, patients' BBT and ambient air SaO2 at the admission were obtained: 25.7% of patients showed BBT above 37.5 °C and 43.5% of them had SaO2 levels under 95%. 307 patients (66.5%) were symptomatic for cough and/or dyspnea. Most patients (66.7%) had at least 3 days of symptoms before getting to the ER. 89.9% of patients had high levels of CRP, 29.0% showed altered WBC count. Past medical history of 462 patients was collected: most frequent comorbidities were hypertension (50.6%), and diabetes mellitus (15%); 372 patients (80.5%) had one or less comorbidity, 90 patients (19.5%) had two or more. Mean number of pre-existing comorbidities was 0.87 (± 0.89). Of all patients, 315 were hospitalized, while 147 were discharged by the physicians. 292 patients (63.2%) met a mild outcome, while for 99 (21.4%) the outcome was moderate and for 71 (15.4%) the outcome was classified as severe. More details about patient characteristics are shown in Table 1.

3.2. ALFABETO performance

The ML proposed behavior (hospitalization vs home care) was confronted with what was done at the emergency room in one of the largest COVID hub in northern Italy by trained clinicians; the behavior at the emergency room held by trained clinicians was considered as the gold standard. 115 Patients were predicted "home" by ALFABETO and discharged to home care by the trained clinicians, 234 patients were predicted "hospital" by ALFABETO and hospitalized by the clinicians, 81 patients were predicted "home" and hospitalized by the clinicians while only 32 patients were predicted "hospital" and discharged to home care by the trained clinicians (Fig. 2).

The performances of ALFABETO calculated with Simple Asymptotic formula and 95% of confidence intervals were: accuracy [71.6, 79.4], AUC [79.5, 86.4], specificity [71.56, 85], recall rate [69.45, 79.11], precision [84, 91.8](see Table 2 for confusion matrix).

Among the patients hospitalized by the clinician 81 patients were predicted to be in "home care" class by ALFABETO. Of these patients 62 (76.5%) did not need ventilation support, 12(14.8%) needed noninvasive ventilation and 7(8.6%) needed invasive ventilation or died; therefore, most misclassified patients (76.54%) showed a favorable/mild clinical course.

Among the patients discharged and predicted as "Hospital" from ALFABETO 30 (93.8%) did not need ventilation, while 1 needed noninvasive ventilation (3.1%) and 1 patient (3.1%) needed invasive ventilation or died.

In the group of the patients admitted to hospital and predicted as "Hospital" from ALFABETO 94 (40.2%) did not need ventilation, while 82 (35.0%) needed noninvasive ventilation and 58 (24.8%) needed invasive ventilation or died.

Among the patients predicted as "Home" from ALFABETO and discharged from a physician 113 (98.3%) did not need ventilation, 1 (0.9%) needed noninvasive ventilation and 1 (0.9%) needed invasive ventilation or died.

A graphical representation of ALFABETO prediction and clinical classification is shown in Fig. 2.

4. Discussion

Since the outbreak of the Covid-19 pandemic, various artificial intelligence software specialized in surveillance, early detection and diagnosis, development and monitoring of treatment, management decision and prevention of COVID-19-related disease have been developed [13,14,15].

Many of the ML and DL software have identified patients affected by COVID-19 focusing on imaging features in X-Rays [16] and CT [17,18], distinguishing covid-19-related from community acquired pneumonia.

Hematologic and biochemical marker abnormalities have been proven to be associated with more severe evolution and mortality in COVID-19 disease [19–22]. Algorithms have been created to predict the evolution of covid-19 in the single patient, integrating laboratoristic with imaging data, in order to optimize the management and the treatment [23–25].

The care and management of COVID-19 patients do not end after an acute infection, but continue in the outpatient setting, in a fraction of patients who report persistent multi-organ (including respiratory, cardio-vascular, neurological systems) symptoms for several months beyond the period of acute infection, the so-called long-COVID. Specifically, Zou et al. proposed AI-assisted chest CT technology to quantitatively measure the extent and degree of lung inflammation in 239 patients who developed pulmonary fibrosis after COVID-19 pneumonia at 30, 60, and 90 days after discharge [26].

ALFABETO performance were in line or slightly worse than those reported in literature for other AI-driven tools [6]; for example, we reported a specificity of 0.78, a recall of 0.74 and a precision of 0.88 while Bararia et al. reported 0.90, 0.77 and 0.74 respectively [27]. Another article with a similar aim as ours was the one from Hao and colleagues that predicted the need for hospitalization with AUC ranges within 86–88% while we reported an AUROC of 83% [21]. An interesting paper, with some difference in the aim (differentiation of common vs severe COVID rather than hospitalization vs home care) from Wei et al. showed AUC 0.93, accuracy 0.91, sensitivity 0.81 and specificity 0.95

Table 1
Patient characteristics.

Mean Age ± SD, years (n = 462)	68.1 ± 15.3 (range 18–100)							
Sex, n (%)(n = 462)	*women*			*men*				
	196 [43]			266 [57]				
BBT, n (%)(n = 485)	*< 37.5 °C*			*≥ 37.5 °C*				
	363 (74.8)			122 (25.2)				
SaO2, n (%)(n = 433)	*≥ 95%*			*< 95%*				
	229 [9,52]			204 [1,47]				
Respiratory Symptoms, n (%)(n = 462)	*no*		*only cough*	*only dyspnea*			*cough + dyspnea*	
	163 (32.6)		60 [12]	203 (40.6)			74 (14.8)	
Days of Symptoms, n (%)(n = 330)	*0–2*		*3–5*	*6–9*			*> 9*	
	106 (32.1)		85 (25.8)	77 (23.3)			62 (18.8)	
Laboratory Tests, n (%)(n = 462)	*WBC*						*CRP*	
	< 4×10³/μL		*> 10×10³/μL*				*> 0.5 ng/dL*	
	81 (16.2)		64 (12.8)				449 (89.8)	
PMH Comorbidities, n (%) (n = 462)	*0*		*1*	*2*			*≥ 3*	
	181 (36.6)		199 (40.3)	80 (16.2)			[34]	
	0.95 ± 0.94 (mean ± SD)			**1 (median)**				
	Hypertension	***DM***	***Dementia***	***Cancer***	***COPD***	***CKD***	***CRF***	***Obesity***
	250 (50.6)	74 [15]	23 (4.7)	30 (6.1)	29 (5.9)	23 (4.7)	9 (1.8)	31 (6.3)
Outcome, n (%) (n = 462)	Mild		Moderate		Severe			
	316 (63.2)		107 (21.4)		77 (15.4)			

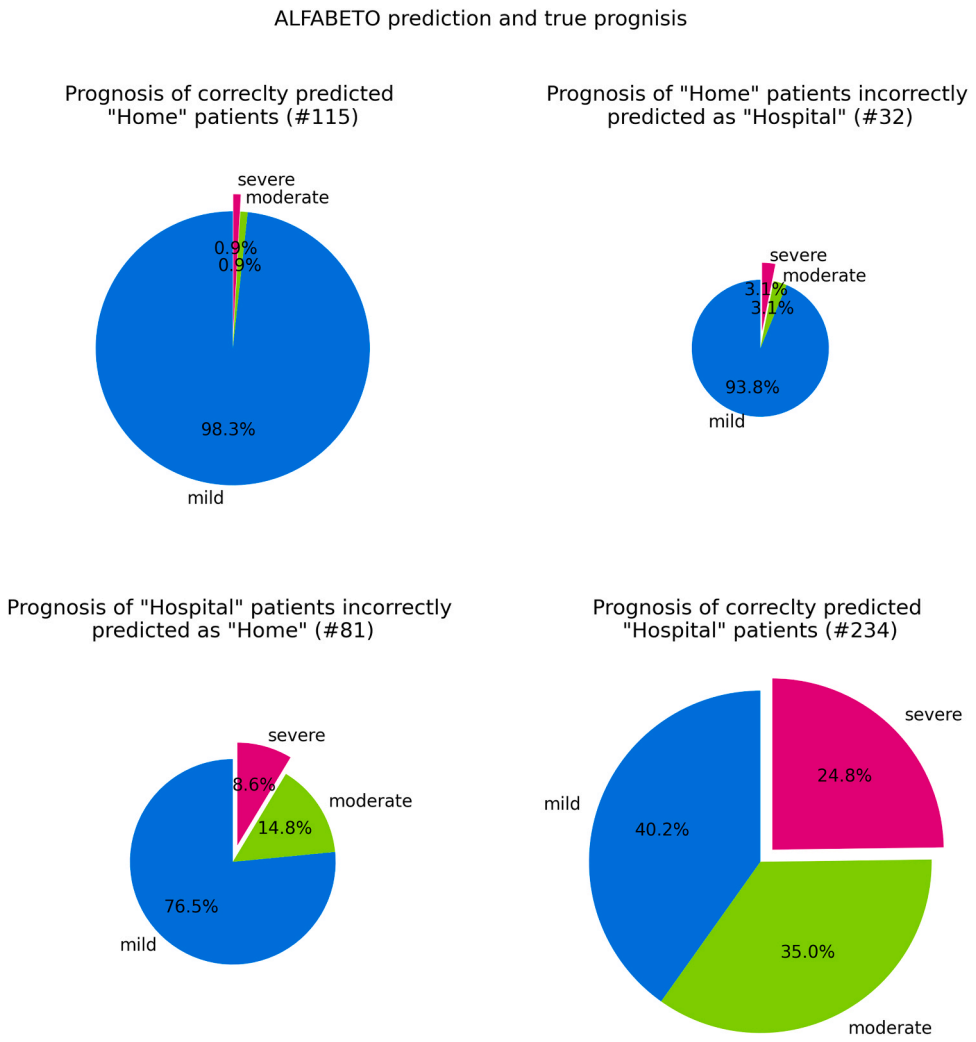

Fig. 2. Prognosis of patients during the third wave. Each pie chart refers to a specific group of patients, identified by the true outcome and ALFABETO predicted outcome. For instance, the lower right pie chart reports the percentage of hospitalized patients correctly predicted as "Hospital" by ALFABETO. The different slices are proportional to the percentage of patients with mild, severe or moderate prognosis in each group.

Table 2
Confusion Matrix of the RF classifier.

	Predicted class	
True Class	Home	Hospital
Home	115	32
Hospital	81	234

[28].

In this context, the quality of the performance may be hampered by dataset shift. Dataset shift arises when patient populations change since training, and it is one of the main causes of reliability lack for the application of ML systems in healthcare [7]. During a pandemic, different sources of shifts arise, from emerging variants associated with increased transmissibility to new treatment protocols defined by healthcare professionals. In our case, we trained the model with data collected during the first wave, and we monitored the performance on third wave patients that were admitted in ER one year after the training of the model. Despite the dataset shift, ALFABETO shows good precision on predicting "Hospital" patients and promising ability to correctly identify "home-care" patients.

For third wave data most of the discrepancies were related to patients predicted home by the AI but hospitalized by the clinician. Discrepancies were 24% of the total with a prevalence for hospitalized patients (patients hospitalized by the ER specialist and classified as "home-care" by ALFABETO) that account for 71% of all discrepancies.

Further analysis aimed at exploring the evolution of these hospitalized patients during the third wave demonstrate a vast majority of patients managed in ordinary wards without ventilation support. Among patients hospitalized by the ER specialist and classified as "home-care" by ALFABETO 3 out of 4 (76.5%) show a mild prognosis while 1 out of 4 (23.4%) show an unfavorable/severe prognosis: 14.8% need noninvasive ventilation and 8.6% need invasive ventilation or deceased. Among patients which were hospitalized by both the AI and the ER specialist figures were drastically different: 2 out of 3 (60%) show an unfavorable/severe prognosis: 35% require noninvasive ventilation and 24.8% need invasive ventilation or died. Only 40.2% show a mild prognosis without the need of ventilation support.

These data allow two considerations. First of all, the evolving landscape of COVID pandemic and its interconnection with the healthcare resources available may reduce the ability of ALFABETO to generalize results over time. Clinicians' decision is influenced by a greater availability of beds/resources during the third wave resulting in more hospitalization in ordinary wards. Another confounding factor could be that all the first wave patients were managed in a tertiary center while during the third wave a hub-spoke network was set in place. Nonetheless it is difficult to couple the discrepancies and rely only on the judgement from the AI; the ability of ER specialist to adapt rapidly to resource

availability is an important ability that AI need to develop as well.

The second consideration is that discrepancies can be used as a strength rather than a weakness. The inability of ALFABETO to adapt to the evolving landscape of the pandemics can be used as a strength since most of patients that show a discrepancy could be managed in spoke centers rather than in a COVID hub and the "discrepancies" between ALFABETO prediction and clinician judgment call may help in selecting them. Patients predicted home care by ALFABETO and hospitalized by the ER specialist, in the less resource constraint environment of the third wave, may benefit from hospitalization but in 3/4 of the cases do not need to be hospitalized in a hub center. On the other hand, patients hospitalized by both ALFABETO and the ER specialist, need to be hospitalized in a hub center since 2/3 of the cases have a moderate/severe outcome.

In this way, the apparent discrepancies, after a clinical review of research results, may result in a further strength, rather than in a weakness, of the interaction between human and AI: if the human and AI agree on hospitalize a Patient it has a high probability of developing a severe disease and need to be admitted to a tertiary center. If AI suggest sending home a Patient that the specialist finds suitable to hospitalize, it can be cost effective to admit it in a spoke center, thus preserving beds in the hub center.

Discrepancies on Patients discharged by ER specialist (and categorized as "hospitalized" by ALFABETO) are only a small percentage of the total and show similar prognosis as those classified as "home care" by both; no significant information can therefore be extracted confronting these two groups.

Our study has limitations influencing testing and validation of classifier results. As already stressed, data are collected in a single center and need a proper external validation. Even if the dataset has a large population, AI frequently needs even larger numbers to be effectively trained and tested. At last, all patients were managed in a tertiary center. Therefore, even if some of them were admitted to ordinary wards, the second opinion and consultation with infectious disease and intensive care specialist were always available; these features may not be available in a real spoke center. This may lead to a survival bias in our population, especially for those patients categorized as mild outcome.

5. Conclusions

The evolving landscape of COVID pandemic and its interconnection with the healthcare resource available may reduce the ability of AI to generalize results over time even though ALFABETO performance were in line or slightly worse than those reported in literature for other AI-driven tools.

Overall, the "discrepancies" emerged between AI suggestion and clinicians actual decision may help in selecting patients suitable to be treated in a spoke rather than in a hub center.

Financial disclosures

The authors have no relevant financial or non-financial interests to disclose.

Funding

MISURA RICERCA COVID19-LINEA2 #RLR12020010362.

Ethical statement

The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This article does not contain any studies with human participants or animals performed by any of the authors requiring local ethic committee approval.

CRediT authorship contribution statement

Conception and design: ChB, LP Administrative support: SC Provision of study materials or patients: MA, MC, AC, LR, CBi, GC, AD, AL Collection and assembly of data: MA, MC, AC, LR, CBi, GC, AD, AL, Data analysis and interpretation: GN, SC FC, GS, AC, SP, RiB, RaB, Manuscript writing: All authors.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Preda Prof Lorenzo reports financial support was provided by Lombardy Region.

References

[1] P. Zhou, X.L. Yang, X.G. Wang, et al., A pneumonia outbreak associated with a new coronavirus of probable bat origin, Nature 579 (7798) (2020) 270–273, https://doi.org/10.1038/s41586-020-2012-7.
[2] S.I. Mallah, O.K. Ghorab, S. Al-Salmi, et al., COVID-19: breaking down a global health crisis, Ann. Clin. Microbiol Antimicrob. 20 (1) (2021) 35, https://doi.org/10.1186/s12941-021-00438-7.
[3] A. Alimadadi, S. Aryal, I. Manandhar, P.B. Munroe, B. Joe, X. Cheng, Artificial intelligence and machine learning to fight COVID-19, Physiol. Genom. 52 (4) (2020) 200–202, https://doi.org/10.1152/physiolgenomics.00029.2020.
[4] G.A. Brat, G.M. Weber, N. Gehlenborg, et al., International electronic health record-derived COVID-19 clinical course profiles: the 4CE consortium, NPJ Digit Med. 3 (2020) 109, https://doi.org/10.1038/s41746-020-00308-0.
[5] S. Subudhi, A. Verma, A.B. Patel, et al., Comparing machine learning algorithms for predicting ICU admission and mortality in COVID-19, May 21, NPJ Digit Med. 4 (1) (2020) 87, https://doi.org/10.1038/s41746-021-00456-x.
[6] M. Roberts, D. Driggs, M. Thorpe, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nat. Mach. Intell. 3 (2020) 199–217, https://doi.org/10.1038/s42256-021-00307-0.
[7] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (1) (2020) 195, https://doi.org/10.1186/s12916-019-1426-2.
[8] Y.S. Malik, S. Sircar, S. Bhat, et al., How artificial intelligence may help the Covid-19 pandemic: Pitfalls and lessons for the future, Rev. Med Virol. 31 (5) (2020) 1–11, https://doi.org/10.1002/rmv.2205.
[9] M. Abdulkareem, S.E. Petersen, The promise of AI in detection, diagnosis, and epidemiology for combating COVID-19: beyond the hype, Front. Artif. Intell. 4 (2021), 652669, https://doi.org/10.3389/frai.2021.652669.
[10] A. Bernheim, X. Mei, M. Huang, et al., Chest CT Findings in Coronavirus Disease-19 (COVID-19): Relationship to Duration of Infection (Jun), Radiology 295 (3) (2020), 200463, https://doi.org/10.1148/radiol.2020200463.
[11] A. Sarica, A. Cerasa, A. Quattrone, Random forest algorithm for the classification of neuroimaging data in alzheimer's disease: a systematic review, Front. Aging Neurosci. 9 (2017) 329, https://doi.org/10.3389/fnagi.2017.00329.
[12] M. Schinkel, K. Paranjape, R.S. Nannan Panday, N. Skyttberg, P.W.B. Nanayakkara, Clinical applications of artificial intelligence in sepsis: a narrative review, Comput. Biol. Med. 115 (2019), 103488, https://doi.org/10.1016/j.compbiomed.2019.103488.
[13] R. Vaishya, M. Javaid, I.H. Khan, A. Haleem, Artificial Intelligence (AI) applications for COVID-19 pandemic, Diabetes Metab. Syndr. 14 (4) (2020) 337–339, https://doi.org/10.1016/j.dsx.2020.04.012.
[14] J. Dong, H. Wu, D. Zhou, et al., Application of big data and artificial intelligence in COVID-19 prevention, diagnosis, treatment and management decisions in China, J. Med Syst. 45 (9) (2021) 84, https://doi.org/10.1007/s10916-021-01757-0.
[15] W. Alsharif, A. Qurashi, Effectiveness of COVID-19 diagnosis and management tools: a review, Radiogr. (Lond.). 27 (2) (2021) 682–687, https://doi.org/10.1016/j.radi.2020.09.010.
[16] R.M. Wehbe, J. Sheng, S. Dutta, et al., DeepCOVID-XR: an artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large U. S. clinical data set, Radiology 299 (1) (2021) E167–E176, https://doi.org/10.1148/radiol.2020203511.
[17] L. Li, L. Qin, Z. Xu, et al., Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy, Radiology 296 (2) (2020) E65–E71, https://doi.org/10.1148/radiol.2020200905.
[18] H.X. Bai, R. Wang, Z. Xiong, et al., Artificial intelligence augmentation of radiologist performance in distinguishing COVID-19 from pneumonia of other origin at chest CT, Radiology 296 (3) (2020) E156–E165, https://doi.org/10.1148/radiol.2020201491.
[19] C.Z. Wang, S.L. Hu, L. Wang, M. Li, H.T. Li, Early risk factors of the exacerbation of coronavirus disease 2019 pneumonia, J. Med. Virol. 92 (11) (2020) 2593–2599, https://doi.org/10.1002/jmv.26071.

[20] G.T. Gerotziafas, T.N. Sergentanis, G. Voiriot, et al., Derivation and validation of a predictive score for disease worsening in patients with COVID-19, Thromb. Haemost. 120 (12) (2020) 1680–1690, https://doi.org/10.1055/s-0040-1716544.
[21] B. Hao, S. Sotudian, T. Wang, et al., Early prediction of level-of-care requirements in patients with COVID-19, Elife 9 (2020), e60519, https://doi.org/10.7554/eLife.60519.
[22] B.M. Henry, M.H.S. de Oliveira, S. Benoit, M. Plebani, G. Lippi, Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis, Clin. Chem. Lab Med. 58 (7) (2020) 1021–1028, https://doi.org/10.1515/cclm-2020-0369.
[23] W. Liang, J. Yao, A. Chen, et al., Early triage of critically ill COVID-19 patients using deep learning, Nat. Commun. 11 (1) (2020) 3543, https://doi.org/10.1038/s41467-020-17280-8.
[24] Y. Gao, G.Y. Cai, W. Fang, et al., Machine learning based early warning system enables accurate mortality risk prediction for COVID-19, Nat. Commun. 11 (1) (2020) 5033, https://doi.org/10.1038/s41467-020-18684-2.
[25] A. Vaid, S. Somani, A.J. Russak, et al., Machine learning to predict mortality and critical events in a cohort of patients With COVID-19 in New York City: model development and validation, J. Med. Internet Res. 22 (11) (2020), e24018, https://doi.org/10.2196/24018.
[26] J.N. Zou, L. Sun, B.R. Wang, et al., The characteristics and evolution of pulmonary fibrosis in COVID-19 patients as assessed by AI-assisted chest HRCT, PLoS One 16 (3) (2021), e0248957, https://doi.org/10.1371/journal.pone.0248957.
[27] A. Bararia, A. Ghosh, C. Bose, et al., Network for subclinical prognostication of COVID 19 Patients from data of thoracic roentgenogram: a feasible alternative screening technology, medRxiv (2020), https://doi.org/10.1101/2020.09.07.20189852.
[28] W. Wei, X.W. Hu, Q. Cheng, Y.M. Zhao, Y.Q. Ge, Identification of common and severe COVID-19: the value of CT texture analysis and correlation with clinical characteristics, Eur. Radiol. 30 (12) (2020) 6788–6796, https://doi.org/10.1007/s00330-020-07012-3.