Review

Anna Carobene*, Frida Milella, Lorenzo Famiglini and Federico Cabitza

How is test laboratory data used and characterised by machine learning models? A systematic review of diagnostic and prognostic models developed for COVID-19 patients using only laboratory data

https://doi.org/10.1515/cclm-2022-0182
Received February 28, 2022; accepted April 22, 2022;
published online May 5, 2022

Abstract: The current gold standard for COVID-19 diagnosis, the rRT-PCR test, is hampered by long turnaround times, probable reagent shortages, high false-negative rates and high prices. As a result, machine learning (ML) methods have recently piqued interest, particularly when applied to digital imagery (X-rays and CT scans). In this review, the literature on ML-based diagnostic and prognostic studies grounded on hematochemical parameters has been considered. By doing so, a gap in the current literature was addressed concerning the application of machine learning to laboratory medicine. Sixty-eight articles have been included that were extracted from the Scopus and PubMed indexes. These studies were marked by a great deal of heterogeneity in terms of the examined laboratory test and clinical parameters, sample size, reference populations, ML algorithms, and validation approaches. The majority of research was found to be hampered by reporting and replicability issues: only four of the surveyed studies provided complete information on analytic procedures (units of measure, analyzing equipment), while 29 provided no information at all. Only 16 studies included independent external validation. In light of these findings, we discuss the importance of closer collaboration between data scientists and medical laboratory professionals in order to correctly characterise the relevant population, select the most appropriate statistical and analytical methods, ensure reproducibility, enable the proper interpretation of the results, and gain actual utility by using machine learning methods in clinical practice.

Keywords: complete blood count (CBC); COVID-19; diagnostic study; laboratory tests; machine learning; prognostic study; SARS-CoV-2.

***Corresponding author: Anna Carobene**, Laboratory Medicine, IRCCS San Raffaele Scientific Institute, Via Olgettina 60, 20132 Milan, Italy, Phone: +39 02 26432850, E-mail: carobene.anna@hsr.it
Frida Milella, IRCCS Istituto Ortopedico Galeazzi, Milan, Italy
Lorenzo Famiglini, DISCo, Università Degli Studi di Milano-Bicocca, Milan, Italy. https://orcid.org/0000-0002-1934-5899
Federico Cabitza, IRCCS Istituto Ortopedico Galeazzi, Milan, Italy; and DISCo, Università Degli Studi di Milano-Bicocca, Milan, Italy

Introduction

Almost 2 years after the COVID-19 pandemic, caused by infection with the novel beta-coronavirus SARS-CoV-2, was declared by the World Health Organization (WHO), there have been over 405 million instances of illness, 5.8 million fatalities globally and over ten billion vaccine doses administered [1].

Early diagnosis is critical in the management of a state of emergency, both for patients affected by COVID-19, whose prognosis may improve because of early therapeutic treatment and for detecting infected asymptomatic subjects [2]. The amplification of viral genomic material (RNA) collected from the upper airways, in particular oro-pharyngeal and/or nasopharyngeal swabs, via rRT-PCR (reverse transcription polymerase chain reaction) is currently the gold-standard method for diagnosing SARS-CoV-2 infection [3]. However, due to the method's sensitivity, the time and cost necessary for the analysis, the need forspecialized equipment, and the associated shortage of reagents at the outbreak of the pandemic, this molecular approach has considerable limits [4, 5].

The need for improved diagnostic capability for SARS-CoV-2 infections, with speedy, accurate, and easily accessible procedures, arose quickly. A common strategy to overcome the COVID-19 pandemic has been the

widespread use of SARS-CoV-2 lateral flow assays, rapid antigen testing of respiratory samples developed by various diagnostic test manufacturers. While some initial reports using lateral flow antigens are promising, at least in terms of specificity, their performance in practice remains controversial [6, 7]. Some researchers concentrated instead their efforts on developing machine learning (ML) models that could aid in the diagnosis and, in some cases, the prognosis of COVID-19 patients [8].

Machine Learning (ML) is a subset of artificial intelligence (AI). It is a word that refers to a number of computational methods that allow a machine to learn from experience and construct algorithms based on data collection, allowing it to complete tasks [9, 10]. In fact, IA/ML approaches are finding a wide range of applications in medicine, as evidenced by the exponential increase of IA/ML/deep learning (DL) publications in recent years (from 203 papers published in 2005 to 12,563 papers indexed on PubMed in 2019 and more than 31,000 in 2021) [11]. Radiology, oncology, and surgery are among the specialties in which IA/ML applications are more prevalent because these are fields in which models are generated with the help of imaging and diagnostic examinations. Laboratory medicine, on the other hand, is still underrepresented, as Ronzio et al. point out [12].

However, the number of publications in this field is increasing [12, 13]. Let us consider the number of articles published in the last 10 years, not just on ML models but, more broadly, on AI studies using lab medicine data (Figure 1). The number of papers has increased at an exponential rate [14]. Herman DS et al. [15], recently published a review of ML systems that are already used in clinical laboratories or have been proposed for application in the recent literature.

Early COVID-19 ML models were based on computed tomography (CT) or chest radiography data and frequently supplemented with conventional molecular diagnostic findings [16, 17].

These studies have yielded promising results, but they have also raised serious concerns, due to the high number of false negatives obtained with chest radiography or the impossibility of using CT for screening due to factors such as high radiation dose, high costs, and a limited number of available instruments [18].

Following that, efforts were focused on the development of ML models based on routine blood test data, beginning with scientific evidence that some blood parameters are significantly altered in COVID-19 patients and can thus serve as good disease markers [19–21]. Laboratory tests have the advantage of saving time and money, as well as being less invasive for the patient, allowing them to be repeated at regular intervals.

Laboratory tests remain a simple, accessible, near real-time, and cost-effective biomarker, reflecting the basic routine blood checks, and they are frequently available in low-resource settings due to their low cost and lack of specific assay equipment. Furthermore, laboratory tests can yield a huge amount of data, so laboratory medicine is an ideal application for ML [22]. ML methods can be used to complement RT-PCR tests to increase the sensitivity of the latter or to provide its assessors with a pre-test probability to calculate NPV and PPV. Additionally, rapid blood test

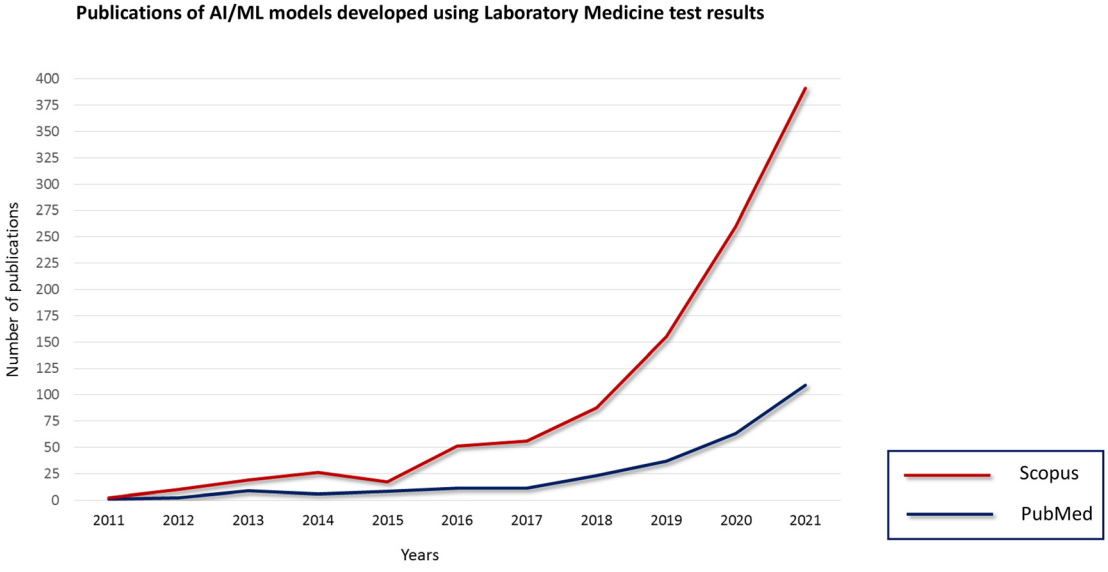

Figure 1: The number of articles connected to artificial intelligence studies based on laboratory medicine data that have been indexed in PubMed (blue line) and Scopus (red line) in the last 11 years (2011–2021).

results may be a valuable clue to early (although inconclusive) identification of COVID-19 patients, leading to better treatment/isolation while waiting for gold standard results [23].

The development of these approaches in the current health emergency is intended to meet the need for valuable models not only for early diagnosis, assisting clinicians in patient management and allowing timely therapeutic intervention but also for prognostic/predictive purposes, i.e., predicting disease progression in order to identify patients at a higher risk of serious adverse events and thus requiring closer monitoring [13, 24, 25].

The goal of this review is to examine the various published studies on ML approaches developed in this area from the start of the pandemic to the present, sorting studies that developed ML models using laboratory data alone or laboratory data accompanied by vital signs/symptoms/comorbidities with a focus on heterogeneity in terms of the input data selection and the accuracy of results obtained from the various models produced. Some variables, such as the importance of the selection and standardization of input data or the external validation of the model, are vital in employing ML in laboratory medicine to describe this heterogeneity.

Methods

A systematic search of the literature was conducted in PubMed [11] and Scopus [26] to identify ML models used as diagnostic and prognostic support tools for COVID-19.

The search was conducted until December 28, 2021, with the results being filtered for the years 2020 and 2021.

- Terms entered into PubMed: Title/Abstract ("blood tests" OR "blood exams" OR "laboratory tests" OR "laboratory exams") AND ("COVID-19" OR "COVID" OR "SARS-CoV-2" OR "coronavirus") AND ("machine learning" OR "deep learning" OR "artificial intelligence").
- Terms entered into Scopus: Title-ABS-KEY ("blood tests" OR "blood exams" OR "laboratory tests" OR "laboratory exams") AND ("covid-19" OR "covid" OR "sars-cov-2" OR "coronavirus") AND ("machine learning" OR "deep learning" OR "artificial intelligence").

Studies that met the following inclusion criteria were included:

- Papers written in English.
- Papers available online in letter, article or conference paper format.
- Papers that presented models built on laboratory data alone or laboratory data accompanied by vital signs/symptoms/comorbidities using ML techniques.
- Papers published in 2020 and 2021 that presented ML models for diagnostic and prognostic purposes.

Aspects of PRISMA (preferred reporting items for systematic reviews and meta-analyses) [27] have been considered in reporting this study.

Results

Literature search results

A PubMed search for the years 2020–2021 yielded 123 publications, and a Scopus search yielded 156. Only 115 publications were reported in both databases, and 98 were eliminated because they did not match the selection criteria, leaving 68 studies to be included in this study (Figure 2).

Of the 68 eligible publications, 34 developed ML models for diagnostic purposes (the studies labeled D1, D2, D3, …, D34), and 34 developed ML models for prognostic tasks (the studies labeled P1–P34).

The complete reference list for the studies chosen is displayed as Supplementary data (Supplementary Table 1).

The first ML diagnostic study applied to laboratory medicine for the management of COVID-19 patients was published in June 2020 (D6), and the first prognostic study was published barely one month later, in July 2020 (P3), according to the literature search (Supplementary Table 1). There are 38 papers published in biomedical journals, 23 in IT-specific publications, and seven IT conference papers (reported solely in Scopus) among the 68 papers chosen for this review (Table 1, Supplementary Table 1). The Materials and Methods sections of the majority of the studies (92%) lack a detailed explanation of the analytical method and/or instruments used in the laboratory analyses data to construct ML models (Table 1). Only six papers mention the instruments used, whereas 39 papers report the unit of measurement. Twelve papers in medical journals do not include any information about the laboratory test that was employed (Table 1). Figures 3–5 describe the characteristics of in the articles reviewed (dataset, purpose of study, features selected, ML models) for diagnostic and prognostic studies, respectively.

Population characteristics

Diagnostic studies

Publications for diagnostic purposes are based on cohorts of subjects with a significant degree of variability in number (ranging from 106 to 115,394). Study D2, for

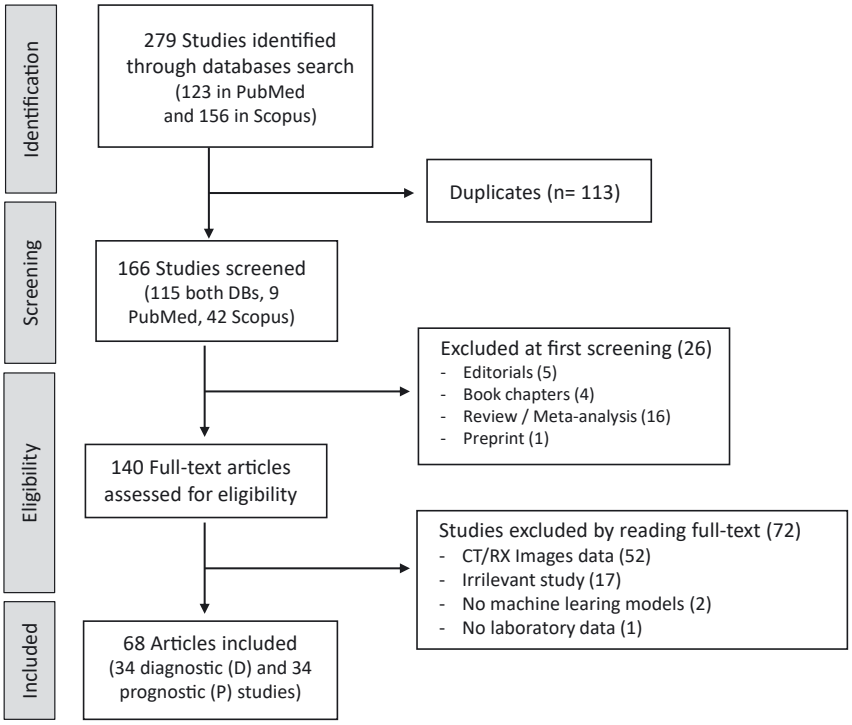

Figure 2: Initial number of PubMed and Scopus publications on COVID-19 patient with diagnostic and prognostic purposes using artificial intelligence approaches (years 2020–2021), number of publications excluded due to selection criteria and final number of papers included in the review.

Table 1: Summary of the information related on laboratory test in the studies selected, differentiated for biomedical journals, IT-specific journal and IT conference papers. For the detailed list of the studies selected, see Supplementary Table 1.

	Biomedical journals (38 papers, 56%)	IT-specific journal (23 papers, 34%)	IT conference papers (7 papers, 10%)	Total (68 papers, 34D and 34P)
Unit of measurement only	20	7	–	27 (40%)
Unit of measurement and reference values	2	2	2	6 (9%)
Unit of measurement and analyzer	2	–	1	3 (4%)
Unit of measurement/analyzer/ analytical principle	2	1	–	3 (4%)
No information	12	13	4	29 (43%)

D, diagnostic study; P, prognostic study.

example, has 171 participants, 24.6% of whom are COVID-19 patients, but the group is adequately defined. In comparison, publication D10, which was published in an IT-focused journal, includes a large cohort (115,394 participants), with only 0.3% of them being COVID-19 positive. The proportion of COVID-19-positive individuals in the population studied is significantly varied (0.3–66.0%, in D10 and D29, respectively). This feature is arguably the most important for genuinely comparing performance across studies.

All diagnostic papers were based on data acquired during the pandemic's initial phase, which was the first seven months of 2020. Four of them were developed in the United States (D1, D3, D5, and D9), five in Italy using different cohorts (D2, D4, D13–14, and D28), two in the United Kingdom (D10, D18), one in Egypt (D29), one in Iran (D26), four in other European countries (Austria (D12), Spain (D16), Slovenia (D17), and Turkey (D29)) and 17 papers (D6–D8, D11, D15, D19, D21–D25, D27, D30–D34) are all related to the same Brazilian cohort and dataset [28] (Figure 3, up panel).

Prognostic studies

Research for prognostic purposes is often based on smaller cohorts of patients than those included in diagnostic studies, and the cohorts are very variable (87 (P4) to 64,733 (P24)). Data for these research projects was obtained in the

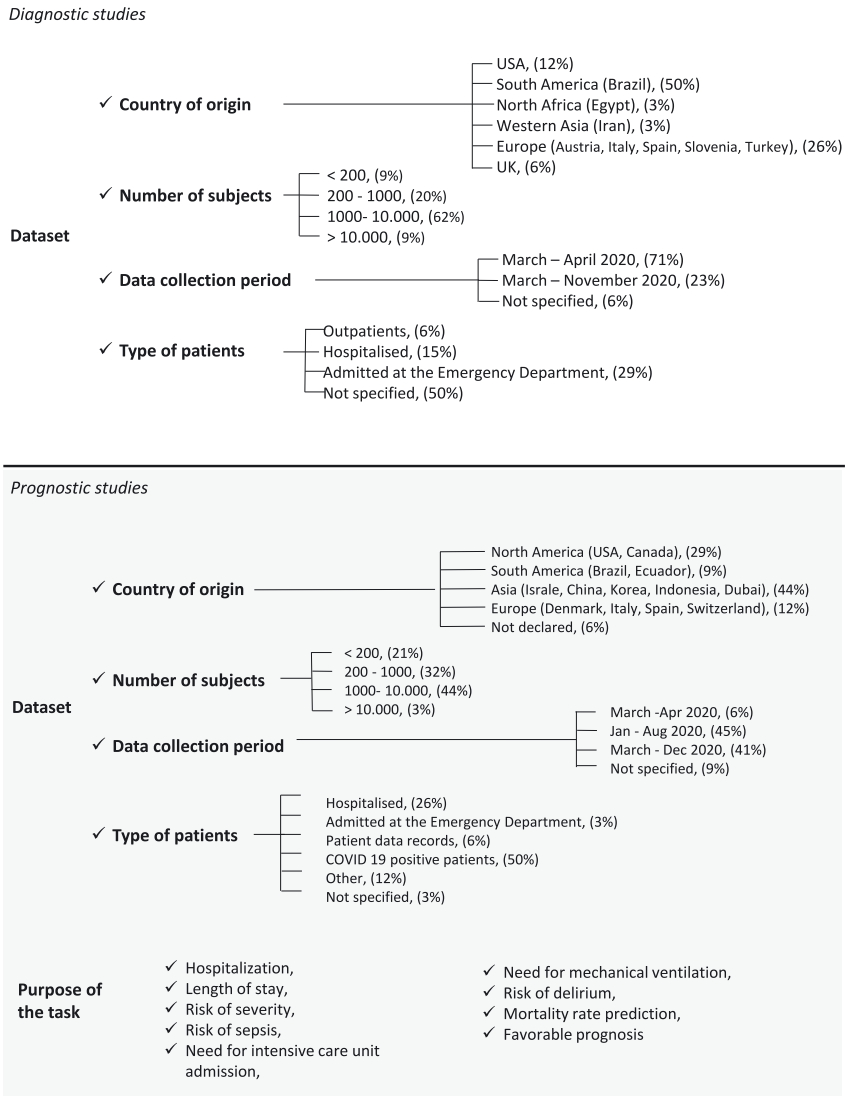

Figure 3: The top and bottom panels, respectively, offer an overview of the variables contained in the diagnostic (D) and prognostic (P) datasets (number and type of individuals and their origin, data collecting time).
The task's purpose is displayed for P studies.

initial months of the pandemic in different nations, with different pandemic curve tendencies, just as it was for the diagnostic studies, even if three studies did not indicate the period of data collection (Figure 3, bottom panel). Patients were tested for COVID-19 using rRT-PCR, and salivary tests or molecular tests on pharyngeal swabs were also used in one study (P6).

Description of prognostic tasks

Prognostic models, as shown in Figure 3, have been developed for a variety of tasks, including the ability to predict the length of time spent in the hospital, the requirement for admission to an intensive care unit, the risk of severity and/or sepsis and the prediction of mortality or a favorable prognosis.

It is worth noting that not all studies use the same definition of "risk of severity": in some, it is defined as intensive care unit admission, mechanical ventilation and/or death, while in others, it is defined as ventilation or death, as well as critical care unit admission or death.

The P8 and P9 research projects produced ML models to predict ICU admission and mortality for distinct purposes, P19 and P23 predicted mechanical ventilation demand and mortality separately, and P28 predicted the duration of stay and mortality. The P5 research project created models for four separate purposes: hospitalisation, admission to an intensive care unit, the need for mechanical ventilation and mortality.

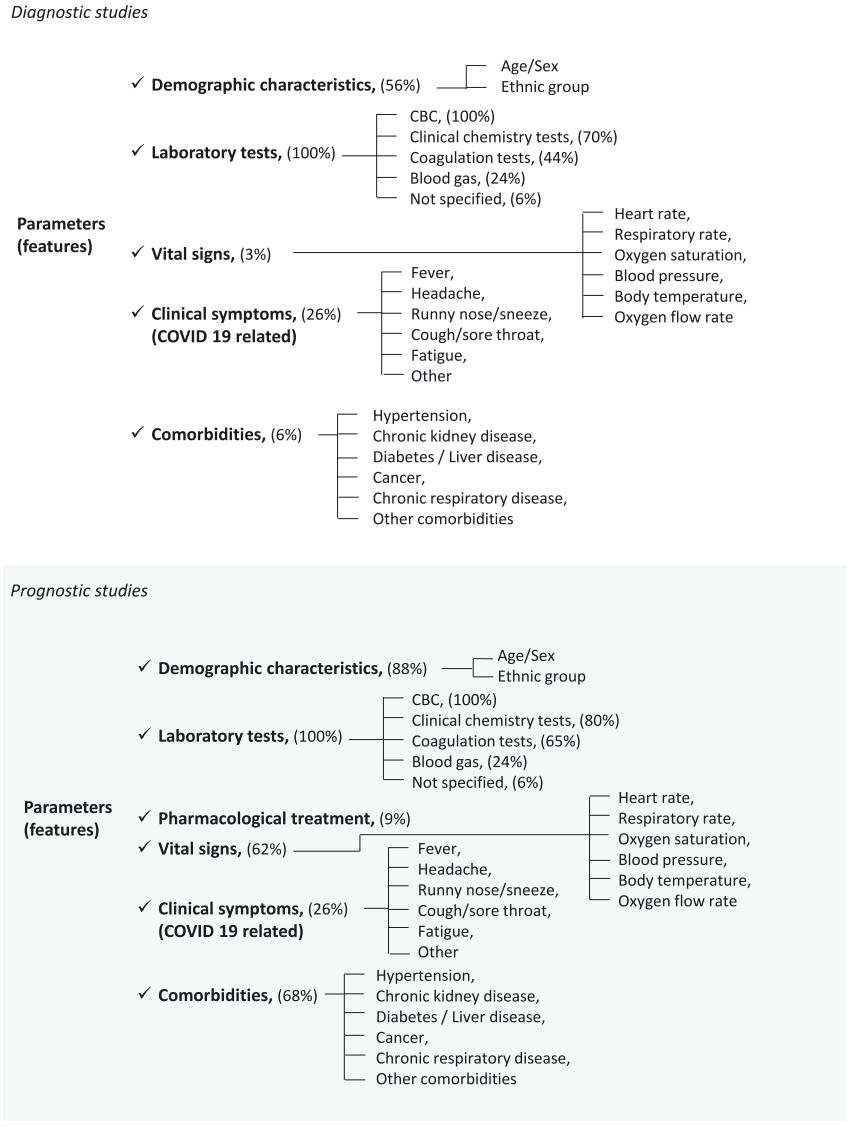

Figure 4: Overview of the variables included in the diagnostic (D) and prognostic (P) studies (demographic characteristics, laboratory tests, pharmacological treatment, vital signs, clinical symptoms, comorbidities).

Parameter (features) descriptions

Diagnostic studies

The number and type of parameters (features) utilised while developing distinct ML models for diagnostic purposes is an important aspect of variability (Figure 4).

All studies evaluate complete blood count (CBC) parameters, 24 additionally consider clinical chemistry values, 15 consider coagulation tests and eight studies also consider blood gas analysis parameters.

Only half of the diagnostic models take demographic data, such as ethnicity, gender and age, into consideration, and only a few studies integrate clinical characteristics, comorbidities and COVID-19-specific symptomatology into the ML model.

Only a few research projects mention the number of features (laboratory parameters) incorporated into the model, without any additional information.

Prognostic studies

CBC and clinical chemistry parameters are frequently included in prognostic models; however, unlike diagnostic models, most research considers other variables. In particular, demographic characteristics are evaluated in 30/34 research projects, whereas comorbidities and vital signs are considered in 23/34 and 21/34 studies, respectively (Figure 4). Only three research projects address pharmacological treatments, while nine studies include symptoms (P15, P19, and P28). Only laboratory test data were used in four studies (P2, P18, P27, and P29).

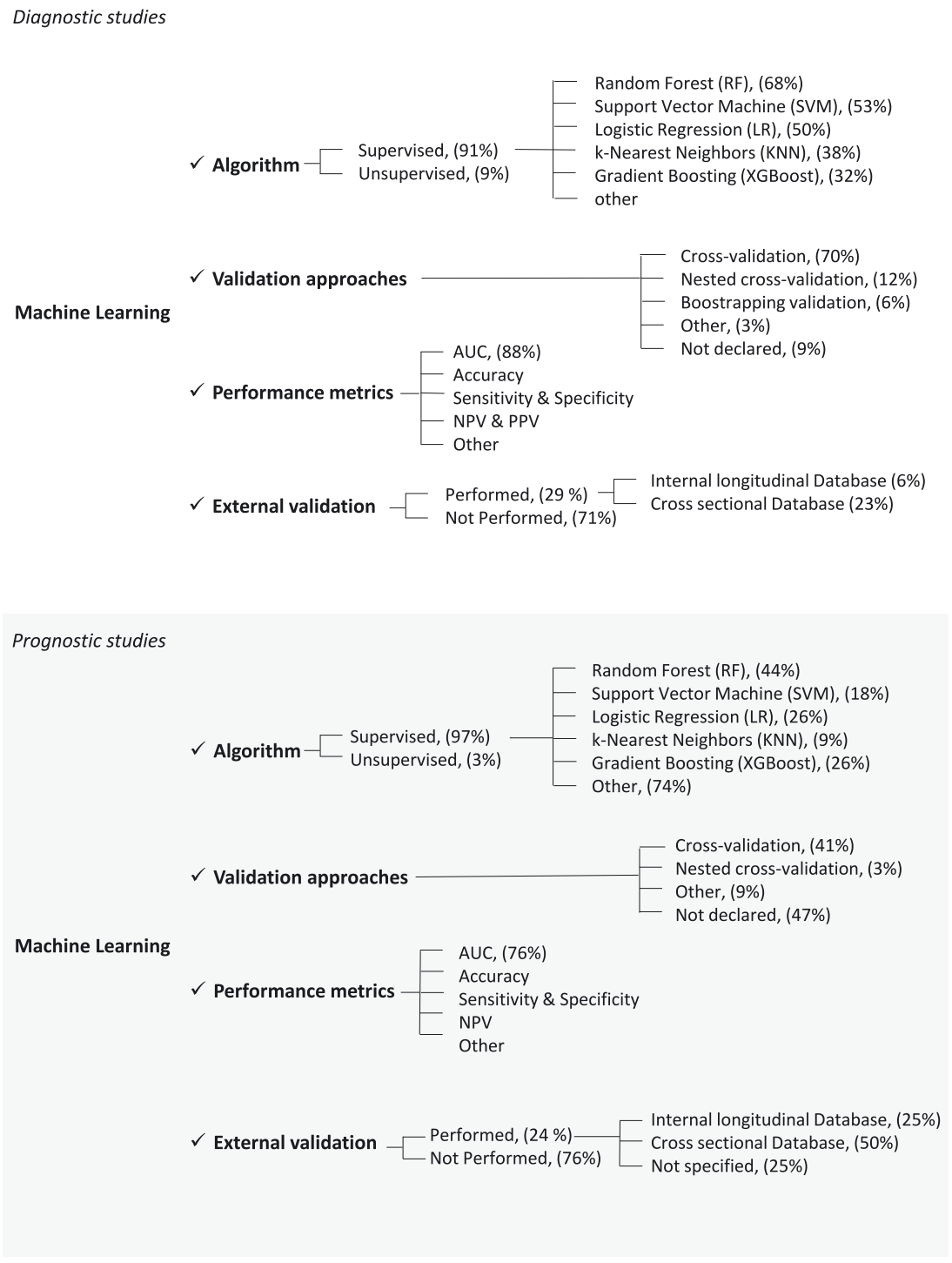

Figure 5: Overview of the variables included in the diagnostic (D) and prognostic (P) studies of the machine learning applications (type of algorithm, mode of data validation, performance metric).

Different sampling methods have only been recorded in a few studies. For example, the P5 trial reported sampling at four separate times: at the time of positive diagnosis, the first 12 h following hospital admission, 12 h before ICU admission, and 12 h after ICU admission.

Coagulation tests were used in 65% of the research, while blood gas test data were used in 24% of the investigations. Inflammatory indicators, including interleukins (ILs), procalcitonin (PCT) and tumor necrosis factor (TNF), were used in a few studies, as well as cardiac

biomarkers such as troponins (TnI/TnT) and brain natriuretic peptide (BNP or proBNP).

ML models description

Diagnostic studies

In the diagnostic papers included for this review, a total of 26 different supervised machine learning models are used, and in the majority of these studies (27 out of 34), the same dataset is investigated with multiple models. Random Forest (RF), support vector machine (SVM), logistic regression (LR), k-nearest neighbours (KNN), and gradient boosting (XGBoost) are the main ML models used in the various investigations (Figure 5).

The principal component analysis (PCA), variational autoencoder (VAE), generative adversarial networks (GAN), restricted Boltzmann machine (RBM) and self-organizing maps (SOM) unsupervised ML models were used in three investigations (D16, D30, and D33).

Prognostic studies

The objective of the prognosis was examined using one ML model in 18 of the 34 studies, while the dataset was investigated by multiple models in the others. Twenty-two supervised ML models are investigated globally in prognostic studies; RF is the most widely used technique, but good accuracy values have also been found with other ML algorithms (Figure 5). Three research projects employ models that incorporate the usage of numerous algorithms (P6, P7, and P20).

Performance metric description

Diagnostic studies

In most research, the outcomes of the ML models are expressed in terms of AUC (Figure 5). The performance of the studies, given in terms of AUC, spans a wide range, from a low of 74% (D12) to a high of 99% (D8, D27, D30). Only ten of the 34 diagnostic investigations disclosed external validation results, two of which used internal longitudinal databases and eight of which used cross-sectional databases.

The support vector machine (SVM) model produced the best accuracy value based on an external validation of the reported data, with an AUC of 98% (D13).

Prognostic studies

The performance measures for most models constructed for prognostic purposes were provided in terms of AUC, although 10/34 projects only reported accuracy, sensitivity and specificity.

Accuracy values of around 80% were reported for the three prognostic purposes listed in the previous paragraph (hospitalization, ICU admission and need for mechanical ventilation), while AUC values of more than 90% were reported in models that were able to predict mortality or an increased risk of developing a severe form of COVID-19 (Figure 5).

Discussion

Sixty-eight publications that applied AI and ML techniques in the COVID-19 context utilizing only laboratory data were chosen from the 166 papers identified by search strings in Scopus and PubMed for our review. The majority of the COVID-19 diagnostic investigations used imaging to create ML models. These studies often produce satisfactory findings, but they are associated with greater expenses, longer timeframes, and more complex patient management, as well as an increased risk of infection inside the hospital and radiology department as compared to models generated with only laboratory data [29].

Several factors should be addressed while developing a meaningful ML model for COVID-19 patients, including the model's objective, the patients to whom it should be applied, the parameters (or features) chosen, the ML algorithm, the performance metrics and the model's validation. It is important to note that such evaluation includes both clinical and laboratory factors, as well as mathematical and statistical aspects.

Only or primarily considering computational aspects can result in models that are not particularly useful or usable. To overcome this, several professions should collaborate in the construction, validation and application of the ML model. The many papers reviewed here will be discussed in light of the aforementioned factors.

Patient selection heterogeneity

The papers reviewed show a great deal of variation in patient selection. The majority of the diagnostic models were developed using emergency department (ED) patient groups, with inpatients and outpatients being included in the remaining investigations. It is worth noting that the

patients' origins were not specified in some studies (D8, D11–12, D15, D20–25, D27, D30–D34).

In several papers, the ML model has been evaluated in cohorts of unselected patients, with the only inclusion criteria being the availability of the molecular swab result; some biochemical tests and, sometimes, age, sex or ethnicity (D1, D3, D7–10). This inaccuracy in patient characterization is common in studies that used the same Brazilian dataset that is freely available online [28]. Inclusion criteria shape the training distribution, and hence affect its representativeness with respect to the target population from which instances will be drawn upon which the machine will produce new prediction/classifications. Therefore, inclusion criteria can affect prospective (actual) accuracy (or better yet robustness) in that they limit the range of types of clinical cases (i.e., instances or phenotypes) about which the machine will be capable to exhibit a certain performance. In other words, inclusion (and exclusion criteria) in the collection of the training dataset represent sources of bias to take into account to expect a specific actual accuracy on the new cases after training.

In a few studies, the authors may assert that they incorporated some clinical parameters into their study, but they do not provide any of them (D5, D12).

In certain cases, only a partial description (the existence of COVID-19-like symptoms) (D3, D8, D14–15, D17, D21–22, D25–26) is provided, and/or some exclusion criteria, such as the presence of specific comorbidities, are also mentioned (D2). Finally, in certain research, subjects' clinical presentation was taken into account and accurately documented (D4), and/or the pattern was analysed in distinct subgroups (the entire sample or the subgroup of asymptomatic vs. symptomatic patients) (D4).

Even the clinical characteristics of a COVID-19 patient, such as clinical presentation, prevalence by age (which has changed over time), the introduction of new therapies (e.g., hyper-immune plasma) and biochemical and instrumental tests, show extreme heterogeneity. From this perspective, a model constructed using a dataset from a specific subgroup that achieved high diagnostic accuracy in one clinical environment may perform poorly in another. "The Importance of Being External" is the provocative title of a recent study. According to the authors, there is currently a gap in the literature regarding how to evaluate external validation results and, hence, assess the robustness of ML models [30]. Via the use of eight external validation CBC datasets collected across three continents during different pandemic waves, authors reported that the correlation between accuracy and similarity should serve as a warning sign that reproducing good performance across very heterogeneous settings can be overambitious and unrealistic [30].

These considerations are essential in the case of predictive models.

The predictive models examined here include a great deal of variation in the population's inclusion criteria and the study's prognostic end-point.

Mortality (at various time points) (P2–3, P5, P7), a severe form of COVID-19 characterized by respiratory failure requiring mechanical ventilation and/or admission to the intensive care unit (P1, P4–5), the risk of sepsis (P30) or delirium (P13), the likelihood of being hospitalized and length of stay (P28) or a favorable prognosis (P6) were all considered as end-points (Figure 3). Comparisons among models are challenging due to the variability of end-point selection.

How the ML models use medical laboratory data

The selection of laboratory tests is a second crucial part of applying the ML model to a patient with a suspected (diagnostic models) or certain (predictive models) diagnosis of COVID-19.

It is important to remember that, while rRT-PCR is considered the gold-standard diagnostic approach, it has limits in terms of sensitivity and specificity, and some additional criteria (e.g., CT and/or clinical presentation) are frequently used to correctly classify patients [23]. As a result, the analytical procedure for COVID-19 diagnosis utilized in the various studies with which the ML model is evaluated is undoubtedly another source of variability.

The laboratory professional should play a vital role in the selection of laboratory tests to be employed in ML models. Although the ML model can theoretically highlight a previously unknown link between a marker and pathology, the features selection should be interested in, or at least include, variables that have already been reported in the scientific literature as associated or altered in the various stages of COVID-19 disease.

A vast number of laboratory tests have been reported to be changed, with putative relevance for monitoring, stratification and prognosis. Specifically, there have been hematological (leukocytosis, lymphopenia, neutrophilia, anemia, thrombocytopenia), biochemistry (hypoalbuminemia, elevated lactate dehydrogenase, aspartate aminotransferase, alanine aminotransferase, total bilirubin, creatinine, troponin, C-reactive protein), infection indicator (interleukins and procalcitonin) and coagulation (increased D-Dimer and prothrombin time)

alterations [18–21, 31, 32]. However, during COVID-19, rapid changes in these parameters are possible, which could be especially problematic for the ML model and its application. The importance of specific laboratory tests may thus change not only during different stages of the disease but also during different eras of the pandemic [14, 30].

The selection of lab tests, similar to patient selection, has a significant degree of variety. There are ML models that use only CBC parameters (D1–2, 6–7, D13–15, D18, D20, D28, P18, P20, P25, P33), others use CBC and clinical chemistry and others also use coagulation tests and/or blood gas tests (Figures 3 and 4).

In addition to laboratory tests, prediction models took into account additional factors, such as clinical history, comorbidities and pharmacological therapy. There are parsimonious models with a small number of variables (less than 10 in D1, D29, P2 and P18) and, more typically, models with a large number of predictors (more than 50 in D4, D10, D21–25 and D27 and P3, P12, P14–15 and P30–31).

The selection of parameters is typically contentious, and it does not appear to be done in consultation with a laboratory medicine expert. For example, a diagnostic model that includes a marker that cannot be requested urgently and has little or no diagnostic significance (such as urea, which was included in various studies, or cholinesterase, included in D12), as well as one that has particularly long analytical times or is excessively expensive, is not very applicable. Curiously, a prognostic study (P27) based on 28 laboratory parameters (including CBC, coagulation and clinical chemistry tests) detected Urea as the most important feature to predict the mortality for patients with COVID positivity. Similarly, a diagnostic or predictive model containing a marker that is rarely found in a hospital laboratory would be useless (i.e., glucose-6-phosphate dehydrogenase included in P28, ILs included in P30 (IL1, IL6, IL8, IL10) and BNP, proBNP, or even atrial natriuretic peptide (ANP) included in P28). Some papers included generic "troponin" (P8, P16, P23) or a generic "bilirubin" (D10) without stating which one (total, direct or indirect bilirubin) it is.

An overview of the laboratory tests, summarised in macro-categories as hematological clinical chemistry and coagulation tests, included in the diagnostic (D) and prognostic (P) studies shown in Figure 6. To the heterogeneity of the numbers of tests used by the different studies, the variety of the expression of the data is also added. For example, WBC differential count data are expressed as percentages, absolute counts or both or in an undeclared unit. Similarly, prothrombin time (PT) is expressed in s, in INR, in terms of percent or in both in s and INR, and in more than 40% of the papers, the unit is not specified (Figure 6, Table 1).

Many factors can influence the representativeness of datasets obtained from hemato-chemical parameters for assessing the robustness of a model: differences in testing equipment (concepts of harmonization), reference ranges/ethnic variability, disease manifestations/phenotypic variability and how humans react to contextual factors (biological variation) make the reference population incredibly vast and diverse; thus, very different datasets can be obtained [30, 33]. This could explain why ML based on lab data has not taken off as it has in other medical fields [13].

The analytical aspect is, in reality, another significant factor to consider (unit of measurement, reference interval, analytical method, instrumentation, traceability). Consider, for example, the non-specificity of Jaffe methods as compared to enzymatic methods in measuring serum creatinine, a parameter used in the majority of the studies examined, as well as the impact that the standardisation process has had on the quality of its measurement [34, 35]. It is clear that, at least for some laboratory parameters, information about analytical methods is essential. In this regard, a mini-review dedicated to D Dimer outcomes in COVID-19 patients is clearly interesting. In this mini-review, there are misunderstandings in terms of how the data are reported, resulting in significant misrepresentation [36]. In our review, more than 90% of the papers selected do not provide sufficient information to adequately characterise laboratory data, and more than 40% do not declare the units of measurement (Table 1).

The use of only CBC data in algorithms developed via ML for diagnostic and prognostic purposes seems to be more appropriate, especially given the robustness of the model based on its reproducibility, rather than the practicability and economic issues [23, 37–40]. In fact, compared to other hemato-chemical parameters, CBC data are characterized by limited within-subjects and between-subjects biological variation [41–43], which supports the reproducibility of results for the same patient at different times, as well as negligible analytical variation [44], which ensures the reproducibility of the same data across laboratories, equipment and heterogeneous populations.

Most diagnostic and predictive models do not offer information on analytic methods and/or instruments, with the exception of a handful (D4, D13, D20, D28, P20) (Table 1). Reviewing the cited papers gives the impression that the authors did not rationally select laboratory parameters but, rather, developed the various models simply by using all of the data available in the management systems of individual healthcare facilities.

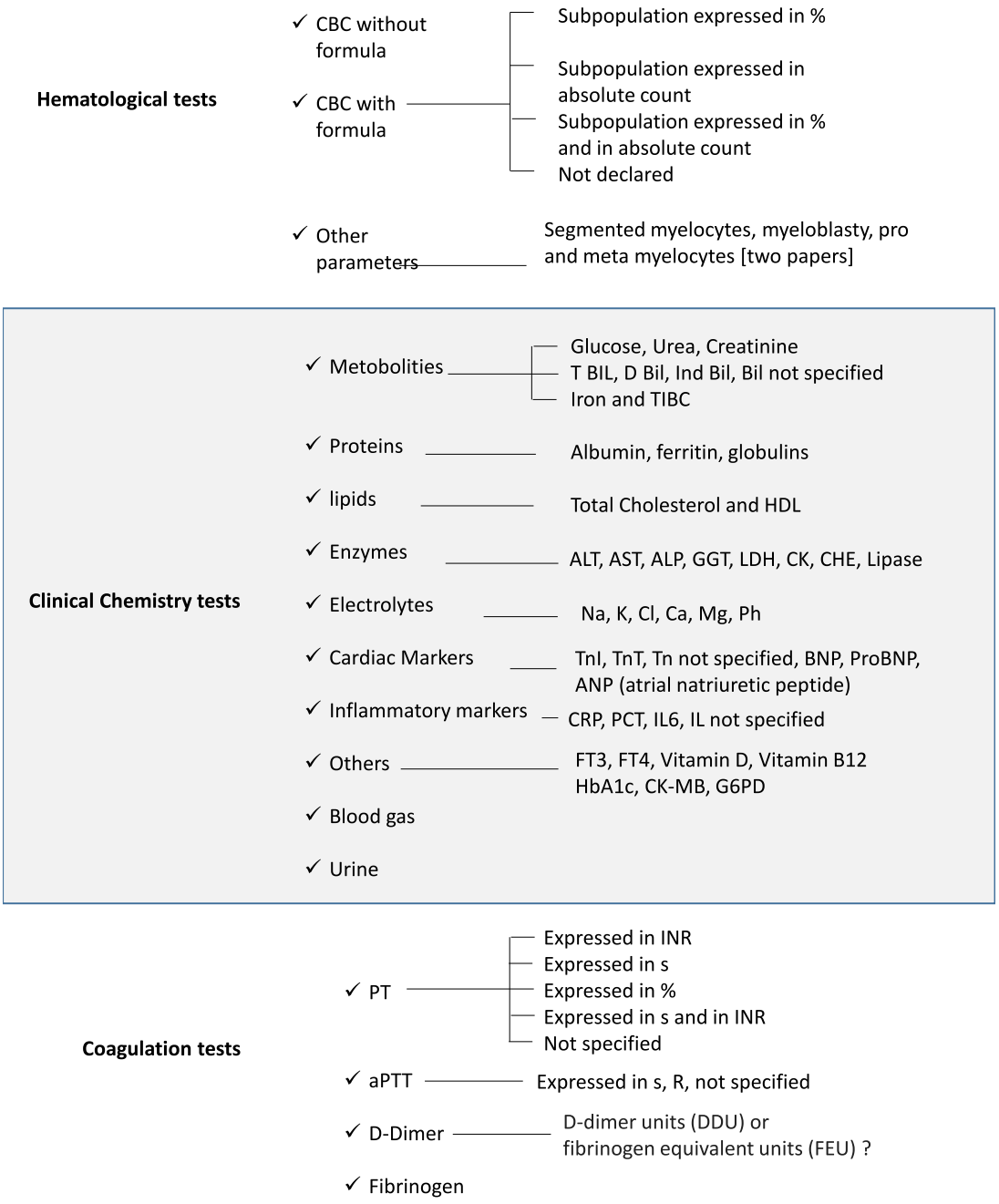

Figure 6: Overview of the laboratory tests included in the diagnostic (D) and prognostic (P) studies.

Due to the large expansion, in recent years, of research using AI techniques, including ML, and the resulting poor reproducibility of the models created [45], it has become imperative to draft standards that could improve the quality of AI studies in the medical area. The MINIMAR (MINimum Information for Medical AI Reporting) guidelines, which were recently published, were established with this goal in mind [46]. MINIMAR includes a minimum list of information that should be included in publications, but it makes no mention of how to record laboratory test

findings, which is crucial to a study's reproducibility, with this being, ironically, the guidelines' goal. Haymond S et al., in an opinion just published [47], summarize practices that should be applied in the development, reporting and review of ML applications without any indication of how clinical lab data should be characterised.

It is no surprise that recommendations for studies involving laboratory medicine results [48–51] always include a request for a thorough description and characterisation of the methods employed to generate them (instrument used, analytical principle, unit of measurement, method optimization, generation, specificity). The analytical processes should measure the same quantity in order to achieve comparable and reliable results, and a description of the analytical approach used to gather laboratory data is required to identify the measurand precisely [52]. Even if the various challenges linked to standardisation and harmonisation and, consequently, the reliability and reproducibility of laboratory results, are not addressed here, as laboratory experts, we are also aware that, if these factors are not considered, the potential for laboratory data to be misused is extremely significant.

ML model types and their validation

The majority of the algorithms utilised are supervised algorithms, with only three diagnostic studies and one prognostic study (D16, D30, D33, P21) developing ML models based on unsupervised methods.

In a nutshell, supervised algorithms attempt to create a model from "labelled" training data, which can then be used to generate predictions about future data. In the unsupervised instance, on the other hand, only input data are provided, and the goal is to model the underlying structure or distribution of the data and discover unknown patterns [53].

However, it is worth noting the number of distinct algorithms (of varying complexity levels) that were used (Figure 5). Specifically, RF (23 out of 34 diagnostic models, 15 out of 34 predictive models) and LR (17/34 and 12/34, respectively) are the most commonly employed supervised algorithms (Figure 5).

The validation process employed in the various research projects is a second point to consider. Split-sample validation, that is, hold-out (by splitting the data into training and test), K-fold cross-validation (which is more resilient than the former), K-fold nested/repeated CV, and bootstrapping approaches for obtaining confidence intervals, is the most commonly used technique. These methods have a variety of benefits and drawbacks [54, 55], and they are examples of internal validation, which is particularly valuable in ensuring model stability. External validation, on the other hand, permits measuring a model's reproducibility and generalisability to the timing and space of data collection by using a dataset that is not the same as that used to train the model. Internal-external validation, a mixture of the two techniques, is also conceivable, and especially effective in situations in which the amount of accessible data is restricted. The authors reported external validation with one or more validation datasets for one-third of the diagnostic models studied (Figure 5). Another important factor to consider is the size of the dataset employed, which influences the validation algorithm and procedure used, as well as the model's correctness. The sample sizes for the diagnostic models range from 170 (D2) to 115,394 (D10), while the predictive models have sample sizes ranging from 87 (P4) to 64,733 (P24).

Model implementation

Two other factors are critical to the model's application: the dataset's representativeness and availability, as well as the algorithm's explainability, robustness and fairness. It is also crucial to make the models publicly available by developing online tools that allow users to evaluate the algorithms that have been developed.

The availability of all of the data used to create the model enables a conclusive evaluation of the model's validity and the verification of its applicability. This is especially important in the context of Big Data, machine learning techniques and other scenarios characterised by great computational complexity: it is a necessary component of real reproducible science [56]. To do this, the original data, source code and detailed descriptions of all analytical methods utilised are required. Many authors are currently unwilling to disclose their data for a variety of reasons; nonetheless, such sharing, as well as data openness in general, is a necessary condition for science to progress on a solid foundation. Although only one research project (D10) made its source code available (at https://github.com/andrewsoltan/CURIAL-manuscript), the datasets for some others are publicly available (D4, D8, D14–15, D21, D23–24, D27, D30, D33–34; P10–12, P15, P20, P26, P29, P32).

The availability of IT tools that allow the model to be implemented by the end-user (the physician) is a second essential factor. These tools can be realised as apps that are directly integrated into the management systems of specific operating units or the laboratory's middleware. Once

the laboratory findings are obtained, this will have the advantage of automating the compilation of input fields, lowering both reaction time and the rate of input errors. In this regard, it is worth noting that some of the diagnostic and predictive models under consideration have been applied in real-world applications that can be found on the Internet (D4, D14, D17, P10, P26).

Conclusions

Some pertinent conclusions can be drawn in light of what has been covered thus far:

- In terms of the type of patients, their number, the laboratory parameters chosen, the algorithms utilised and their validation, the research projects considered are marked by extraordinary heterogeneity.
- The use of diagnostic and prognostic models necessitates an accurate patient description and an appropriate characterization of the lab test used in terms of unit, instrument and analytical principle, as well as the statistical and analytical methods used (and, if possible, the publication of the dataset and source code). Additionally, the availability of a web application or, potentially, a computer tool incorporated into hospital management systems, facilitates the model's implementation and usability.
- Effective collaboration between data scientists and clinicians is essential in obtaining clinically useful models; however, laboratory medicine professionals should gain a thorough understanding of Big Data and machine learning algorithms to understand their potential and correctly interpret their results.

The authors of "A Brief Guide to Medical Professionals in the Age of Artificial Intelligence" [57] expect that clinicians will devote time to learning the fundamentals of these new technologies so that they can assess clinical study opportunities. The clinical laboratory community's involvement is, in fact, critical to ensuring that laboratory data are sufficiently available and incorporated conscientiously into strong, safe and clinically successful ML-aided diagnoses [15].

Research funding: None declared
Author contribution: All authors have accepted responsibility for the entire content of this manuscript and approved its submission.
Competing interests: Authors state no conflict of interest.
Informed consent: Not applicable.
Ethical approval: Not applicable.

References

1. WHO. WHO Coronavirus (COVID-19) Dashboard; 2020. Available from: https://covid19.who.int/ [Accessed Feb 2022].
2. Oran DP, Topol EJ. Prevalence of asymptomatic SARS-CoV-2 infection: a narrative review. Ann Intern Med 2020;173:362–7.
3. Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DK, et al. Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Euro Surveill 2020;25:2000045.
4. Dinnes J, Deeks JJ, Adriano A, Berhane S, Davenport C, Dittrich S, et al. Cochrane COVID-19 diagnostic test accuracy group. rapid, point-of-care antigen and molecular-based tests for diagnosis of SARS-CoV-2 infection. Cochrane Database Syst Rev 2021;3: CD013705.
5. Campagner A, Carobene A, Cabitza F. External validation of machine learning models for COVID-19 detection based on complete blood count. Health Inf Sci Syst 2021;9:37.
6. Ogawa T, Fukumori T, Nishihara Y, Sekine T, Okuda N, Nishimura T, et al. Another false-positive problem for a SARS-CoV-2 antigen test in Japan. J Clin Virol 2020;131:104612.
7. Kretschmer A, Kossow A, Grüne B, Schildgen O, Mathes T, Schildgen V. False positive rapid antigen tests for SARS-CoV-2 in the real-world and their economic burden. J Infect 2022;84: 248–88.
8. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:1328.
9. Naugler C, Church DL. Automation and artificial intelligence in the clinical laboratory. Crit Rev Clin Lab Sci 2019;56:98–110.
10. Razavian N, Major VJ, Sudarshan M, Burk-Rafel J, Stella P, Randhawa H, et al. A validated, real-time prediction model for favorable outcomes in hospitalized COVID-19 patients. NPJ Digit Med 2020;3:130.
11. PubMed. National Library of Medicine; 2021. Available from: https://pubmed.ncbi.nlm.nih.gov/ [Accessed Feb 2022].
12. Ronzio L, Cabitza F, Barbaro A, Banfi G. Has the flood entered the basement? a systematic literature review about machine learning in laboratory medicine. Diagnostics 2021;11:372.
13. Cabitza F, Banfi G. Machine learning in laboratory medicine: waiting for the flood? Clin Chem Lab Med 2018;56:516–24.
14. Carobene A, Sabetta A, Monteverde E, Locatelli M, Banfi G, Di Resta C, et al. Machine Learning based on laboratory medicine test results in diagnosis and prognosis for COVID-19 patients: a systematic review. Biochim Clin 2021;348:64.
15. Herman DS, Rhoads DD, Schulz WL, Durant TJS. Artificial intelligence and mapping a new direction in laboratory medicine: a review. Clin Chem 2021;67:1466–82.
16. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ 2020;369:1328.
17. Mei X, Lee HC, Diao KY, Huang M, Lin B, Liu C, et al. Artificial intelligence-enabled rapid diagnosis of patients with COVID-19. Nat Med 2020;26:1224–8.
18. Ulhaq A, Born J, Khan A, Gomes DPS, Chakraborty S, Paul M. COVID-19 control by computer vision approaches: a survey. IEEE Access 2020;8:179437–56.
19. Fan BE, Chong VCL, Chan SSW, Lim GH, Lim KGE, Tan GB, et al. Hematologic parameters in patients with COVID-19 infection. Am J Hematol 2020;95:131–4.

20. Ferrari D, Seveso A, Sabetta E, Ceriotti D, Carobene A, Banfi G, et al. Role of time-normalized laboratory findings in predicting COVID-19 outcome. Diagnosis (Berl) 2020;7:387–94.
21. Ferrari D, Cabitza F, Carobene A, Locatelli M. Routine blood tests as an active surveillance to monitor COVID-19 prevalence. a retrospective study. Acta Biomed 2020;91:e2020009.
22. Vidali M. I big data e la medicina di laboratorio. Biochim Clin 2021;45:13–4.
23. Cabitza F, Campagner A, Ferrari D, Di Resta C, Ceriotti D, Sabetta E, et al. Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. Clin Chem Lab Med 2021;59:421–31.
24. Deo RC. Machine learning in medicine. Circulation 2015;132: 1920–30.
25. Luo Y, Szolovits P, Dighe AS, Baron JM. Using machine learning to predict laboratory test results. Am J Clin Pathol 2016;145:778–88.
26. Scopus. Available from: https://www.scopus.com/search/form.uri?display=basic#basic [Accessed Feb 2022].
27. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med 2009;6: e1000100.
28. Kaggle ED; 2020. Available from: https://www.kaggle.com/einsteindata4u/covi [Accessed Dec 2021].
29. Carobene A, Campagner A, Sulejmani A, Leoni V, Seghezzi M, Buoro S, et al. Identification of Sars-CoV-2 positivity using machine learning methods on complete blood count data: external validation of state-of-the-art models. Biochim Clin 2021; 45:281–9.
30. Cabitza F, Campagner A, Soares F, García de Guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Progr Biomed 2021;208:106288.
31. Lippi G, Plebani M. Laboratory abnormalities in patients with COVID-2019 infection. Clin Chem Lab Med 2020;58:1131–4.
32. Rodriguez-Morales AJ, Cardona-Ospina JA, Gutiérrez-Ocampo E, Villamizar-Peña R, Holguin-Rivera Y, Escalera-Antezana JP, et al. Clinical, laboratory and imaging features of COVID-19: a systematic review and meta-analysis. Trav Med Infect Dis 2020; 34:101623.
33. Badrick T, Banfi G, Bietenbeck A, Cervinski MA, Loh TP, Sikaris K. Machine learning for clinical chemists. Clin Chem 2019;65: 1350–6.
34. Carobene A, Ceriotti F, Infusino I, Frusciante E, Panteghini M. Evaluation of the impact of standardization process on the quality of serum creatinine determination in Italian laboratories. Clin Chim Acta 2014;427:100–6.
35. Paroni R, Fermo I, Cighetti G, Ferrero CA, Carobene A, Ceriotti F. Creatinine determination in serum by capillary electrophoresis. Electrophoresis 2004;25:463–8.
36. Favaloro EJ, Thachil J. Reporting of D-dimer data in COVID-19: some confusion and potential for misinformation. Clin Chem Lab Med 2020;58:1191–9.
37. Formica V, Minieri M, Bernardini S, Ciotti M, D'Agostini C, Roselli M, et al. Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2. Clin Med 2020;20:114–9.
38. Avila E, Kahmann A, Alho C, Dorn M. Hemogram data as a tool for decision-making in COVID-19 management: applications to resource scarcity scenarios. PeerJ 2020;8:e9482.
39. Banerjee A, Ray S, Vorselaars B, Kitson J, Mamalakis M, Weeks S, et al. Use of machine learning and artificial intelligence to predict SARS-CoV-2 infection from full blood counts in a population. Int Immunopharm 2020;86:106705.
40. Famiglini L, Bini G, Carobene A, Campagner A, Cabitza F. Prediction of ICU admission for COVID-19 patients: a machine learning approach based on complete blood count data. In: IEEE 34th international symposium on computer-based medical systems, CBMS; 2021:160–5 pp.
41. Buoro S, Carobene A, Seghezzi M, Manenti B, Dominoni P, Pacioni A, et al. Short- and medium-term biological variation estimates of red blood cell and reticulocyte parameters in healthy subjects. Clin Chem Lab Med 2018;56:954–63.
42. Coskun A, Braga F, Carobene A, Tejedor Ganduxe X, Aarsand AK, Fernández-Calle P, et al. Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters. Clin Chem Lab Med 2019;58:25–32.
43. Buoro S, Carobene A, Seghezzi M, Manenti B, Pacioni A, Ceriotti F, et al. Short- and medium-term biological variation estimates of leukocytes extended to differential count and morphology-structural parameters (cell population data) in blood samples obtained from healthy people. Clin Chim Acta 2017;473:147–56.
44. Vidali M, Carobene A, Apassiti Esposito S, Napolitano G, Caracciolo A, Seghezzi M, et al. Standardization and harmonization in hematology: instrument alignment, quality control materials, and commutability issues. Int J Lab Hematol 2021;43:364–71.
45. National Academies of Sciences. Engineering, and medicine, policy and global affairs, committee on science, engineering, medicine, and public policy. In Reproducibility and replicability in science. Washington (DC): National Academies Press (US); 2019.
46. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (Minimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. J Am Med Inf Assoc 2020;27:2011–5.
47. Haymond S, Master SR. How can we ensure reproducibility and clinical translation of machine learning applications in laboratory medicine? Clin Chem 2022;68:392–5.
48. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Standards for reporting of diagnostic accuracy group. the STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Croat Med J 2003;44: 639–50.
49. Bartlett WA, Braga F, Carobene A, Coşkun A, Prusa R, Fernandez-Calle P, et al. A checklist for critical appraisal of studies of biological variation. Clin Chem Lab Med 2015;53:879–85.
50. Zhang GM, Guo XX, Zhu BL, Zhang GM, Bai SM, Wang HJ, et al. Establishing reference intervals of aspartate aminotransferase-to-platelet ratio index for apparently healthy elderly. Clin Lab 2016;62:135–40.
51. Aarsand AK, Røraas T, Fernandez-Calle P, Ricos C, Díaz-Garzón J, Jonker N, et al. The biological variation data critical appraisal checklist: a standard for evaluating studies on biological variation. Clin Chem 2018;64:501–14.

52. Vesper HW, Myers GL, Miller WG. Current practices and challenges in the standardization and harmonization of clinical laboratory tests. Am J Clin Nutr 2016;104:907–12.
53. Carobene A, Campagner A, Uccheddu C, Banfi G, Vidali M, Cabitza F. The multicenter European biological variation study (EuBIVAS): a new glance provided by the principal component analysis (PCA), a machine learning unsupervised algorithms, based on the basic metabolic panel linked measurands. Clin Chem Lab Med 2022;60:556–8.
54. Steyerberg EW, Harrell FE Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol 2016;69:245–7.
55. Steyerberg EW, Harrell FE Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. J Clin Epidemiol 2001;54:774–81.
56. Vidali M. La scienza riproducibile. Biochim Clin 2020;44: 386–96.
57. Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digit Med 2020;3:126.

Supplementary Material: The online version of this article offers supplementary material (https://doi.org/10.1515/cclm-2022-0182).