The Importance of Being External

Methodological insights for the external validation of machine learning models in medicine

Federico Cabitza [a,1,*], Andrea Campagner[a,1], Felipe Soares[b], Luis Garcia de Guadiana Romualdo[c], Feyissa Challa[d], Adela Sulejmani[e], Michela Seghezzi[f], Anna Carobene[g]

[a] *University of Milano-Bicocca, Viale Sarca 336, 20126, Milano, Italy*
[b] *Computer Science Department, University of Sheffield, Sheffield, UK*
[c] *Laboratory Medicine Department, Hospital Universitario Santa Lucía, Cartagena, Spain*
[d] *National Reference Laboratory for Clinical Chemistry, Ethiopian Public Health Institute, Addis Ababa, Ethiopia*
[e] *Laboratorio di chimica clinica, Ospedale di Desio e Monza, ASST-Monza, Dipartimento di medicina e chirurgia, Università di Milano-Bicocca, Monza, Italy*
[f] *Laboratorio di chimica clinica, Ospedale Papa Giovanni XXIII, Bergamo, Italy*
[g] *IRCCS Ospedale San Raffaele, Via Olgettina, 60, 20132, Milano, Italy*

Abstract

Background and Objective Medical machine learning (ML) models tend to perform better on data from the same cohort than on new data, often due to overfitting, or co-variate shifts. For these reasons, *external validation* (EV) is a necessary practice in the evaluation of medical ML. However, there is still a gap in the literature on how to interpret EV results and hence assess the robustness of ML models.

Methods We fill this gap by proposing a *meta-validation* method, to assess the soundness of EV procedures. In doing so, we complement the usual way to assess EV with an assessment in terms of the dataset cardinality, as well as with a novel method that considers the similarity of the EV dataset with respect to the training set. We then investigate how the notions of cardinality and similarity can be used to inform on the reliability of a validation procedure, by integrating them into two summative data visualizations.

Results We illustrate our methodology by applying it to the validation of a state-of-the-art COVID-19 diagnostic model on 8 EV sets, collected across 3 different continents. The model performance was moderately impacted by data similarity (Pearson $\rho = .38$, $p < .001$). In the EV, the validated model reported good AUC (average: .84), acceptable calibration (average: .17) and utility (average: .50). The validation datasets were adequate in terms of dataset cardinality and similarity, thus suggesting the soundness of the results. We also provide

*Corresponding author: e-mail: federico.cabitza@unimib.it.
[1] These authors contributed equally.

a qualitative guideline to evaluate the reliability of validation procedures, and we discuss about the importance of proper external validation in light of the obtained results.
Conclusions In this paper, we propose a novel, lean methodology to: 1) study how the similarity between training and validation sets impacts on the generalizability of a ML model; 2) assess the soundness of EV evaluations along three complementary performance dimensions: discrimination, utility and calibration; 3) draw conclusions on the robustness of the model under validation. We applied this methodology to a state-of-the-art model for the diagnosis of COVID-19 from routine blood tests, and showed how to interpret the results in light of the presented framework.

Keywords: Medical Machine Learning, Validation, Dataset Similarity, Dataset Cardinality, COVID-19

---

1. Introduction

The validation of machine learning (ML) classification models represents one of the most important, and yet most critical, steps in the development of this class of decision support tools [61]. In this respect, to "validate" means to provide evidence that the model is *valid*, that is it will properly work with new data that the model has never examined or processed before.

In the specialist literature, validation of ML models is often intended and performed as *internal* validation [70]: this refer to validation protocols, including, e.g. hold-out, bootstrap or cross- validation, that attempt to estimate the performance of the ML models by partitioning the whole training dataset into multiple smaller datasets, and by testing the model, trained on one part of the original dataset, on a different, usually smaller, part [33, 61]. This class of approaches has prompted the researchers to focus on an important aspect to assess the soundness of the validation procedure, namely the size of the dataset used, or its *cardinality* [49]. We argue, however, that sample cardinality alone is not sufficient for understanding the reliability of a validation procedure, and must be complemented with an equally important aspect, which is often completely overlooked: dataset *similarity*.

While internal validation procedures are widely used, especially for their convenience, they are not considered sufficiently conservative in so-called critical settings, like the medical one [5, 51, 58, 61]. In these settings, ML models must be *robust*, that is capable to reliably work also in contexts that may be more or less subtly different from the one from which the training data has been obtained [30, 59, 66]. This is sometimes called the requirement for *cross-site transportability* [57]. This requirement is due either because the model must be deployed in a different setting, as it is the case of medical ML models that are to be deployed in multiple hospitals or countries [25]; or because the distribution of the underlying phenomenon of interest and predictive variables may change over time (a phenomenon known as concept drift [31]), making the original setting of model deployment a new setting for any practical purpose.

Furthermore, the results of internal validation procedures are sometimes incorporated in the development of ML models, for example as a means to perform model selection [16, 58]: as a consequence, ML models are often not capable to generalize well beyond their training distribution and may be at risk of *data leakage* and *overfitting*, thus leading to highly inflated estimates of their prospective accuracy [16, 38].

Therefore, in critical settings, *external* validation has been advocated as necessary [5, 19, 35, 59]. External data, in this case, refers to a set of new data points that come from other cohorts, facilities, or repositories other than the data used for model creation. Most of the times, performance observed on external datasets is significantly poorer than performance appraised on original datasets (e.g., [39]), so that the following question should be addressed any time researchers develop a ML model: will its performance be reproduced consistently across different sites [54]?

In what follows, we will share a general method to assess the soundness of an external validation procedure grounding on the two notions mentioned above, *dataset cardinality* and *dataset similarity*.

To illustrate this method, we will apply it to the case of the COVID-19 diagnosis [12]. In particular, we will report about the challenges that we met in the validation of a state-of-the-art ML model with external datasets coming from across three continents, as well as of the lessons that we learnt in the interpretation of the results. Finally, we will share a set of practical recommendations for the *meta-validation* of external validation procedures (that is their validation), so as to meet the requirements of generalizability and reproducibility that diagnostic and prognostic ML models must guarantee in daily (clinical) practice [4].

2. Methods

In this Section, we describe our methodological contribution for the assessment of the soundness of external validation procedures. This contribution combines recent metrics and formulas, and integrates them together to get a tool for the qualitative (also visual) assessment of the validity of an external validation procedure. For this reason, we see our proposal as a lean method for *meta-validation*.

As said above, this method integrates two different sets of metrics. One set of metrics is aimed at evaluating the *minimum dataset cardinality* that is necessary to extract meaningful estimates of accuracy from a validation procedure; these metrics are based on the formulas proposed in [8, 49]. By contrast, the other metric was proposed by us in [13] to get an estimate of the (multi-variate) similarity between two datasets: here we apply it to compare the (external) validation set and the training set. In what follows, we briefly present both metrics and then we discuss how we used them to build up a meta-validation procedure.

2.1. Dataset Cardinality Evaluation

As anticipated in the Introduction, high-quality ML studies require that external validation is performed, to assess the capability of an existing model to generalize across different sites, and thus to provide a reliable and reproducible estimate of its performance. To this aim, the sample size of the external validation datasets is an important criterion [2, 49, 64]. Indeed, small sample sizes can result in imprecise or overly optimistic performance estimates [5].

For these reasons, several studies have developed metrics to evaluate the adequacy of a sample size to perform sound external validations, that is to determine a Minimum Sample Size (MSS) sufficient to guarantee the generalizability of the results of an external validation procedure. Traditionally, formulas for computing the MSS have been based on rules-of-thumb approaches [20, 62, 64]. However, recent studies [55] have raised awareness on several limitations of rules-of-thumb formulas, due to the fact that these techniques do not take into account the variance of the ML model with respect to the validation population, or its expected performance.

For these reasons, in recent years, researchers have focused on the development of more precise evaluation formulas, based on either simulation approaches [55] or concentration bounds [45, 49]. Both these methods provide a quantitative indication of a minimum sample size which is sufficient to ensure that the performance estimation, on an external validation set having that sample size, is representative of the true performance of the ML model with high probability.

In what follows, we describe the formulas and computation methods of the MSS for the AUC, the Standardized Net Benefit and the Brier Score. These three performance metrics are targeted at different, but important, dimensions of model performance, namely *discrimination*, *utility* and *calibration* (respectively). In particular, in regard to the AUC and the Standardized Net Benefit, we adopt the formulas for MSS evaluation proposed by Riley et al. [49]. On the other hand, for the Brier score, we adopt the formula for MSS evaluation proposed by Bradley et al. [8]. We chose to adopt these techniques because, compared to other existing proposals, they requires minimal assumptions and provide data-dependent bounds.

In regard to the AUC, let C be the AUC of the ML model on the external validation set, Φ be the proportion of the positive class in the external validation set, and $SE(C)$ the targeted value of the standard error for C. Then, fixed a value for $SE(C)$ (which determines the size of the confidence interval associated with the MSS), the MSS for the AUC can be computed according to the following formula:

$$MSS(AUC) = \min n \in \mathbb{N} \text{ s.t.}$$
$$SE(C) \leq \sqrt{\frac{C(1-C)\left(1 + (n/2-1)\frac{1-C}{2-C} + \frac{n/2-1}{1+C}\right)}{n^2\Phi(1-\Phi)}} \qquad (1)$$

In regard to the Standardized Net Benefit, let sNB_τ be the Standardized Net Benefit of the ML model on the external validation set (at fixed probability threshold τ), Φ be the proportion of the positive class in the external validation set, *Sens* (resp. *Spec*) the sensitivity (resp. specificity) of the ML model on the external validation dataset, $SE(sNB_\tau)$ the targeted valued of the standard error for sNB_τ, and $w = \frac{(1-\phi)\tau}{\phi(1-\tau)}$. Then, fixed a value for $SE(sNB_\tau)$ (which determines the size of the confidence interval associated with the MSS), the MSS for the Standardized Net Benefit can be computed according to the following formula:

$$MSS(sNB) = \frac{\frac{Sens(1-Sens)}{\phi} + \frac{w^2 Spec(1-Spec)}{1-\phi} + \frac{w^2(1-Spec)^2}{phi(1-\phi)}}{SE(sNB_\tau)^2} \tag{2}$$

Lastly, in regard to the Brier Score, let B be the Brier Score of the ML model on the external validation set, p_i the predicted probability score for the i-th case in the external validation set, y_i the true target class for the i-th case in the external validation set, n the cardinality of the external validation set, ϵ be the targeted size of the confidence interval associated with the MSS. Then, the $SE(B)$ and the MSS can be estimated from the external validation dataset as:

$$SE(B) = \frac{1}{n}\sum_i p_i^4 - \frac{4}{n}\sum_i p_i^3 y_i + \frac{6}{n}\sum_i p_i^2 y_i - \frac{4}{n}\sum_i p_i y_i + \sum_i p_i - B^2 \tag{3}$$

$$MSS(B) = \left(\frac{2 \cdot t_\epsilon \cdot SE(B)}{0.05}\right)^2 \tag{4}$$

where t_ϵ is the ϵ-critical value for a Student's t distribution with $n-1$ degrees of freedom.

As mentioned above, the main advantage of the adopted MSS formulas lies in their being distribution-free and in their capability to take into account data-dependent information, namely by relying on the observed performance values in the validation datasets. By contrast, their main disadvantage lies in the inability to take into account either model complexity (e.g., regularized models are more robust than data interpolators, and thus they could require a smaller MSS), or feature dimensionality (e.g., due to the *curse of dimensionality*, a higher MSS would be required when the number of features is large). Future work should be aimed at the development of MSS formulas that can better take into account these contextual information.

2.2. Dataset Similarity Metric

The relationship between data similarity and generalization properties of ML models was first proposed by Bousquet et al. [7]: The authors observed that datasets found to be strongly dissimilar likely originated from different distributions. As a consequence, information about similarity could provide

useful indications to understand why a ML model performs poorly on a validation set [36], and how to perform domain adaptation successfully [48]. To this aim, different metrics to measure data similarity have since been proposed in the literature: Bousquet et al. [7] proposed a metric (called Data Agreement Criterion - DAC), based on the Kullback-Leibler divergence, which has been widely adopted to evaluate the similarity between distributions [63]; Schat et al. [53] proposed an adjustment on the DAC metric (called Data Representativeness Criterion - DRC), and studied the relationship between data similarity and generalization performance; Cabitza et al. [13] proposed a different metric, called *Degree of Correspondance* (denoted as Ψ), based on a multi-variate statistical testing procedure. Notably, both the metric proposed by Bousquet et al. [7] and Schat et al. [53] are based on a parametric approach. Thus, they cannot be easily applied when expert knowledge about appropriate distributions for the data and phenomenon under study is scarce, or when it is not possible to reliably estimate the parameters of the generating distributions [1, 6]. For this reason, for our meta-validation procedure we adopt the *Degree of Correspondance* (Ψ) proposed in [13]. Since this latter technique is non-parametric and distribution-free it is not subject to the above limitations. The procedure to compute the *Degree of Correspondance* is reported in Algorithm 1 as a reference. Intuitively, the Ψ among the two datasets is defined as the p-value for a multi-variate topological test for equality of distributions. As regards the ∂ deviation metrics in Algorithm 1, we employed the Maximum Mean Discrepancy metrics proposed in [28], as the authors of [13] previously showed this version of the *Degree of Correspondance* to be more robust than others. A Python implementation of this algorithm is available at `https://github.com/AndreaCampagner/qualiMLpy/` and a sandbox is provisionally running at `https://reprdeg-test.herokuapp.com/`.

Algorithm 1 The algorithm procedure to compute the similarity between the two dataset T and V, using the *Degree of Correspondance* (Ψ).

procedure $\Psi(T, V$: datasets, d: distance, ∂ deviation metrics)
  $d_T = \{d(t, t') : t, t' \in T\}$
  For each $v \in V$, find $t_v \in T$, nearest neighbor of v in T
  $T_{|V} = \{t \in T : \nexists v \in V.t = t_v\} \cup V$
  $d_{T_{|V}} = \{d(t, t') : t, t' \in T_{|V}\}$
  $\delta = \partial(d_T, d_{T_{|V}})$
  Compute $\Psi = Pr(\delta' \geq \delta)$ using a permutation procedure
  return Ψ
end procedure

As previously mentioned, the main advantage of this metric lies in its non-parametric, distribution-free and multivariate nature, as well as in its computational efficiency. The main disadvantage lies in it being inscrutable in regard to the extent specific features could influence the score. For this reason, extensions of this metric would consider the predictivity of the features, and further re-

search could assess the extent few significantly different feature distributions in each dataset (acting as a sort of outliers) could impact the score and therefore the reliability of any conclusion about the practical relevance of the heterogeneity between the datasets at hand.

2.3. Meta-validation method

In this section, we introduce our proposal for a lean meta-validation procedure, which takes into account both the dataset cardinality (as measured through the MSS formula) and the dataset similarity (as measured through the Ψ metric) within a two-step procedure.

Our procedure encompasses both quantitative and qualitative (in particular visual) elements articulated in two different steps. The first step is aimed at getting a first estimate of the *robustness* of the ML model, interpreted in terms of the susceptibility and dependence of its performance on the dis(similarity) between training and (internal) test sets. For this reason, this step does not require an external validation dataset and it can be performed by exploiting a cross validation, or bootstrap, procedure.

In this step, a linear regression should be derived by modelling the relationship between the dataset similarity (measured by means of the degree of correspondence presented above), seen as an explanatory variable, and any balanced performance metric of choice (e.g. Balanced Accuracy, F-score), seen as the dependent variable. In particular, we suggest to used balanced performance metrics to better account for potential imbalances in the target distribution [9]. Obviously, the resulting model makes no pretensions at being general, as it is highly dependent on (among other things) the model architecture, the selected features, the hyper-parameter settings and the task target. That notwithstanding, such a linear model could give developers informative hints about the extent the heterogeneity of unseen data (of similar kind and pertinent to that particular task) is relevant, with respect to the data patterns the model could learn, in terms of impact on the expected performance of the model and, hence, its robustness. These hints regard three different, but related, elements of the linear model, namely the correlation coefficient (r, and its statistical significance); the coefficient of determination (that is, R^2); and the angular coefficient (b). More precisely, the correlation coefficient provides information in regard to the impact of dataset similarity on model generalization: low correlation values imply stronger generalizability, while higher correlation values imply greater impact of dataset heterogeneity on the reproducibility of the model's results. To guide the interpretation of the correlation values, a widely known convention [18] suggests to interpret correlation coefficients (r) lower than .1 as either absent or negligible correlation between data similarity and model performance; between .1 and .3 as weak or low correlation; values between .3 and .5 as moderate correlations; values between .5 and .7 as strong correlations; and above .7 as very strong correlations. The correlation coefficient can also be given an interpretation in terms of the corresponding R^2 value, which can be analytically expressed as $R^2 = r^2$. Consequently, strong correlations are those for which linear regressions models explain at least one quarter of the variation in model performance by variations

in the value of the similarity between datasets (indeed, in these cases the coefficient of determination R^2 is greater than .25). Finally, the relationship between b and r (indeed, $b = r\frac{\sigma_y}{\sigma_x}$) can be used to derive a graphical representation of the "strength" of the relationship between dataset similarity and model performance, according to the diagram represented in Figure 1, which we call the *potential robustness diagram*, a sort of extended scatterplot.

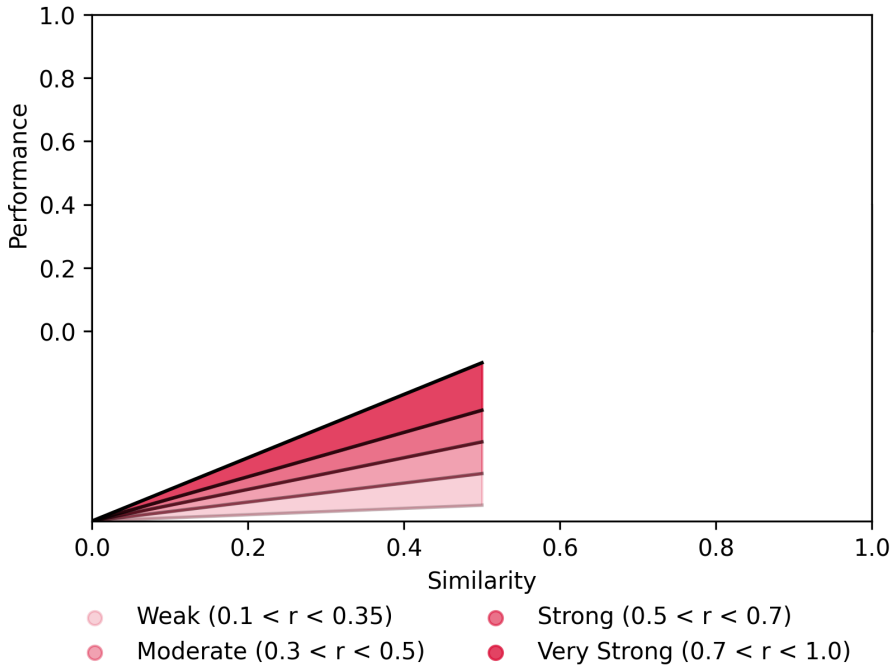

Figure 1: The *potential robustness diagram*, proposed to support the proposed meta-validation methodology at step 1.

In this diagram, the top half should represent a scatterplot of the relationship between dataset similarity and model performance, as obtained through the above mentioned cross-validation, or similar, procedures. The bottom half, on the other hand, represents four "correlation regions", which allow to classify the measured correlation: low angular coefficients could be seen as an indirect indicator of model robustness and tolerance to data variability, as they correspond to either weak (i.e. $r < 0.3$) or absent ($r < 0.1$) correlation. By contrast, high angular coefficients, associated with at least a *moderate* correlation (i.e. $r > 0.3$) would hint at a potentially relevant impact of data variability on the performance and robustness of the ML model. The *potential robustness diagram* allows to compare and interpret actual performance results, as they are produced in the following step of the procedure, in light of the extent they either confirm or challenge this sort of "simulation of generalizability".

The second step of the meta-validation procedure, on the other hand, requires to have at least one external validation dataset (and the more datasets, the better). This step is aimed at assessing the performance of the ML model in light of two different dimensions: dataset similarity (of the external validation dataset, with respect to the training set of the ML model); and dataset cardinality, in terms of sufficient sample size. The performance is evaluated in terms of discrimination, calibration, and utility, three dimensions of equal importance in the comprehensive evaluation of a model quality (although they usually attract different attention at development time, with the former being the most pursued one [62]). The second step requires then to perform a meta-evaluation of the validation procedure, to understand if this latter procedure can be considered conservative and reliable enough. On a practical level, we suggest to perform this evaluation by means of a graphical representation of the above mentioned information, as shown in the diagram in Figures 2, that we call the *external performance diagram*. This diagram allows to depict, for any external validation dataset considered, whether the Minimum Sample Size (MSS) has been achieved (or, possibly, exceeded), in terms of opacity; and three complementary quality dimensions in light of the similarity with respect to the internal datasets: model discrimination power (AUC); model utility (Net Benefit); and model calibration (Brier Score).

In order to maintain consistent naming conventions when describing the relative dataset similarity associated with the degree of correspondance Ψ, the following labels (see Table 1), inspired by the famous nomenclature adopted by Landis and Koch [37], are assigned to the corresponding ranges of the *Degree of Correspondance* (Ψ) and adopted in the proposed diagram.

Ψ	Level of similarity
$<.001$	extremely low
.001-.2	low
.21-.4	slight
.41-.6	moderate
.61-.8	substantial
.81-1	essential

Table 1: Levels of similarity with respect to the value of the *Degree of Correspondance* (Ψ).

Thus, as a rule of thumb, a similarity higher than 60% (i.e., substantial or essential) should make readers wary of the utility of such a validation to tell something about the actual replicability of the model performance. Conversely, good performance exhibited by the model on external datasets that are less than 40% similar with respect to the training set (slight or low similarity) should be considered a reliable test bench in terms of conservative estimates of model performance.

A similar terminology is also adopted in regard to the model performance. In particular, with respect to AUC, values greater than 0.7 are deemed acceptable [41]; while values greater than 0.8, or 0.9, are termed, respectively, good

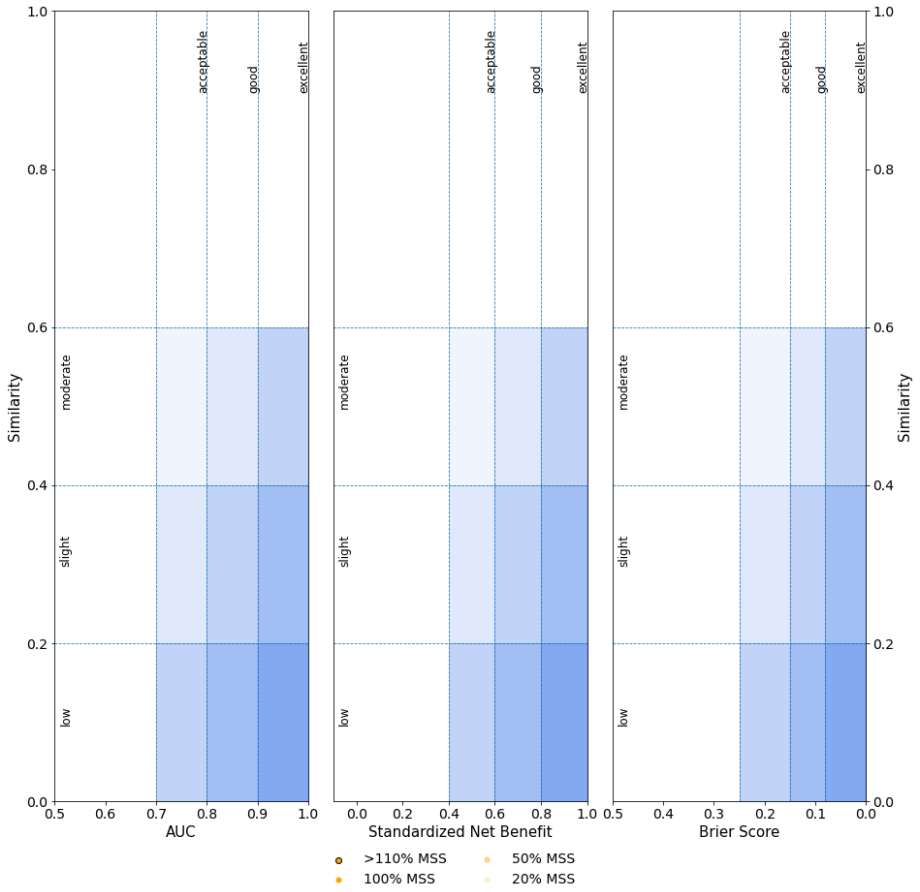

Figure 2: The *external performance diagram*, proposed to support the meta-validation procedure at step 2.

and excellent. Similar thresholds are also adopted for the (Standardized) Net Benefit [52] and the Brier Score [60]. Obviously, other thresholds may be more appropriate for different metrics, such as the Matthews Correlation Coefficient [17] or the Index of Balanced Accuracy [26, 27].

This notions are then represented graphically in the diagram in Figure 2. In particular, in each of the three figures in the diagram, the bottom portion represents the area of low similarity. If the performance of a validation dataset falls into this region, then the validation process can be considered conservative enough; moreover, if a score falls into the right-bottom region then the validation process can be considered as providing a conservative indication of good cross-site transportability.

Figure 2 also provides an evaluation of the external validation procedure in terms of the adequacy of the dataset cardinality, with respect to the Minimum

Sample Size in terms of hue brightness or opacity.

As a rule of thumb, results that are only *acceptable* on datasets that are either *substantially* or *essentially* similar to the training set should be considered *not valid*; conversely, an external validation where at least acceptable AUC scores are observed on at most *slightly similar* external datasets would suggest that the model is valid and robust (with respect to that performance dimension).

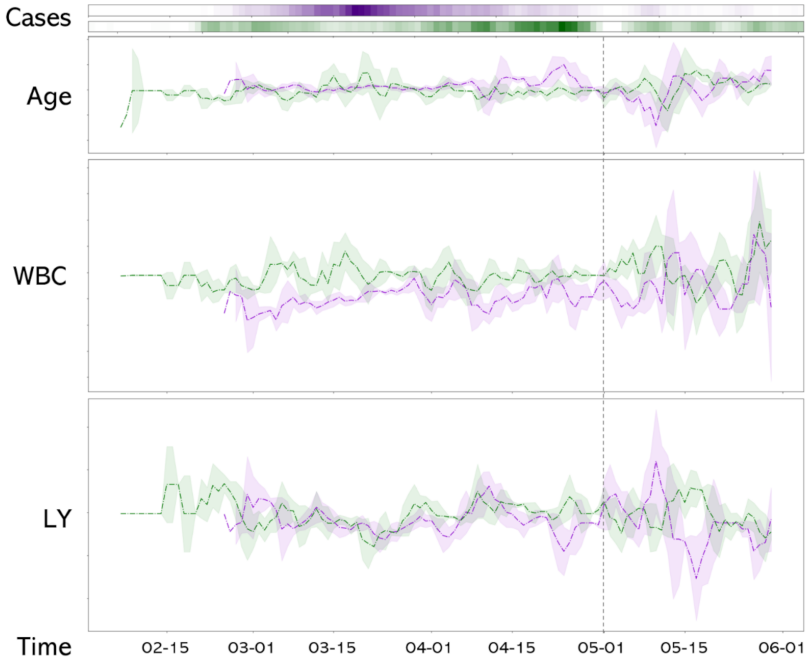

Figure 3: Above: Heat map of the temporal evolution of the rolling 7-day average of the number of the confirmed COVID-19 positive cases (in purple), negative cases (in green) collected in the training dataset: the brighter the color, the higher the number of cases. Below: temporal series the Age, White Blood Count and Lymphocites of the patients admitted at the HSR, with their 95% confidence intervals, chosen for their high predictive value in the diagnostic ML model. The vertical dotted line indicates the 1st of May 2020, when visual inspection allowed to notice some changes in the temporal patterns of the features, while the number of admissions of positive patients was significantly decreasing.

3. Use case: external validation of models supporting COVID-19 diagnosis

In this section, we describe the experimental setting where we applied our methodological approach, the characteristics of the training and validation datasets, and we briefly discuss the results of the validation of a ML model to diagnose COVID-19.

3.1. Related Works on COVID-19 Diagnosis

Since its initial spread in January 2020, the COVID-19 pandemic has so far affected more than 180 million people in 18 months, and caused almost four million deaths worldwide (if not more). For these reasons, different ML models have been applied to aid clinicians in the detection of COVID-19, using different methods, e.g. deep learning model based on imaging data [3, 10, 32, 44, 46] or statistical learning approaches, mainly based on hematochemical parameters [12, 47, 56, 67, 69]. These latter methods, in particular, have been considered particularly interesting for clinical purposes. Indeed, while most imaging-based studies have been found lacking in terms of methodological soundness [50, 68], hematochemical-based models are often more rapid and cost-effective [23], and also more safe, especially when compared to CT procedures [29].

Nonetheless, although the potential of ML methods for COVID-19 detection is high, only a few models have been subjected to external validation [50, 68]. For instance, if we limit ourselves to ML models grounding on hematological data, among tens of publications [68] only the following publications report an external validation procedure to date: [12, 47, 56, 67, 69].

This striking lack of validation studies makes COVID-19 a paradigmatic case, not only for the urgent need of reliable studies and good models through which to improve practice and ultimately health outcomes during health crises, but also for the related lack of reproducibility [50], which has been recently denoted as one of the main challenges to overcome for the real-world adoption of ML-based approaches in medical practice [4, 68].

For this reason, in what follows we will consider the application of the proposed methodological framework to the external validation of one of the few proposals in the literature to have undergone external validation, namely the state-of-the-art ML model proposed in [12].

3.2. Experimental Setting

In what follows, we describe the experimental setting for our study, by which we aim to illustrate the proposed meta-validation methodology. To this purpose, as previously mentioned, we will describe our experience on the external validation of a state-of-the-art COVID-19 diagnostic model [12], based on complete blood count (CBC) data.

This ML model was developed on the basis of a training set encompassing 1736 instances and 21 features, collected upon admission at the emergency departments (ED) of two hospitals in the Milan area (northern Italy) that are 15 km apart: the IRCCS Hospital San Raffaele (HSR), a large teaching hospital of 1,350 beds, serving a catchment area of 39,000 citizens, which was heavily impacted by the first wave of the COVID pandemic; and the IRCCS Istituto Ortopedico Galeazzi (IOG), a teaching hospital specialized in musculoskeletal disorders of 364 beds, serving a catchment area of 15,000 citizens[2].

[2]https://www.dati.lombardia.it/Sanit-/Bacino-di-utenza-delle-strutture-per-ospedale/gbyc-bhps/data

The data were collected between March 5, 2020, and May 26, 2020, that is during the first wave of the COVID-19 pandemic in Northern Italy (see Figure 4). The requirement for the collection phase was to get a training set which was sufficiently balanced and heterogeneous: for this reason, we collected test results from patients admitted in a hospital that was specifically devoted to manage COVID-19 patients (the HSR) with the results of patients who had been mostly admitted to another hospital (the IOG) for other problems (mainly trauma) and were found COVID-positive. Moreover, focusing on the HSR subset, we noticed that the number and case mix of the patients admitted into the ED changed over time, from the first phase of the pandemic (February - April) to the last phase of the first wave (see Figure 3), also reflecting the number of cases at national level (see Figure 4).

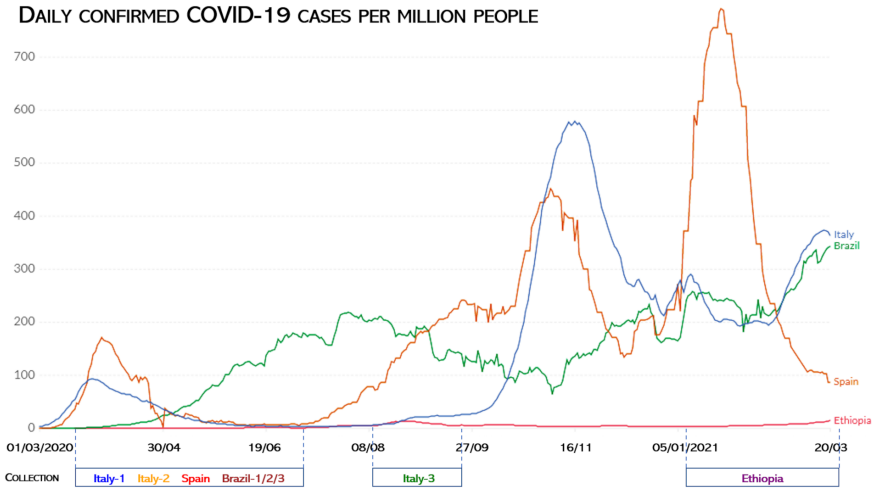

Figure 4: Temporal series of the rolling 7-day average of the number of daily new confirmed COVID-19 cases per million people in the countries where the data (horizontal bars) have been collected for the development and validation of the diagnostic ML model considered in this study. Adapted from ourworldindata.org.

We performed eight external validation procedures on the basis of as many corresponding external datasets, namely:

1. The Italy-1 dataset: This dataset was collected at the Desio Hospital, a general hospital of 383 beds, 25 kilometres due north of Milan, serving a catchment area of approximately 12,000 citizens[1]. This dataset encompasses 337 instances (163 positive, 174 negative) collected in March/April 2020, that is during the first wave of the COVID-19 pandemic in Northern Italy.
2. The Italy-2 dataset: this dataset was collected at the 'Papa Giovanni XXIII' Hospital of Bergamo, a general hospital of 1080 beds, 54 kilometres due east of Milan, serving a catchment area of approximately 28,000

citizens: This dataset encompasses 249 instances (104 positive, 145 negative) collected in March/April 2020, during the first wave of the pandemic.

3. The Italy-3 dataset (from the IRCCS Hospital San Raffaele, that is the same from where the training cases had been collected), which encompasses 224 instances (118 positive, 106 negative) collected in November 2020, that is 5 months later the collection of the training set, during the peak of the second wave of COVID-19 in Italy;
4. The Spain dataset. This dataset was collected at the University Hospital Santa Lucía in Cartagena, a city of 215,000 inhabitants in the Region of Murcia, in Spain: This dataset encompasses 120 instances (78 positive, 42 negative) collected in October 2020, in one of the two hospitals of the above Region.
5. The 3 Brazil datasets: The first dataset, Brazil-1, was collected in the Fleury private clinics thorough Brazil; while the other 2 datasets, Brazil-2 and Brazil-3, com respectively from the Albert Einstein Israelite Hospital, and the Hospital Sírio-Libanês. These latter two hospitals of, respectively 627 and 466 beds, are both located in Sao Paulo (Brazil), with a catchment area of approximately 23 million people, and are considered two of the most important hospitals in Brazil and South America. The 3 Brazil datasets encompass, respectively, 1301 (352 positive, 949 negative), 2335 (375 positive, 1960 negative) and 345 (334 positive, 11 negative) instances, all collected between February 2020 and June 2020, during the first wave of the pandemic in Brazil;
6. The Ethiopia dataset. This dataset was collected at the National Reference Laboratory for Clinical Chemistry, Millenium COVID-19 Treatment and Care Center, of the Ethiopian Public Health Institute in Addis Ababa, and encompasses 400 (200 positive, 200 negative) instances collected between January and March 2021.

The set of predictive features is reported in Table 2, while the characteristics of the datasets are summarized in Table 3. The four most predictive quantitative features [12] and the prevalence of suspect symptoms are reported in Figures 5 and 6.

In regard to the characteristics of the ML model, we evaluated a pipelined model encompassing: a missing data imputation step (based on K-Nearest Neighbors); a data standardization step; and a classification model based on Support Vector Machine classifier with RBF kernel (see [12] for details on model development and optimization). As previously mentioned, this model was trained on a dataset encompassing 1736 instances and 21 features, collected at the HSR and IOG hospitals, and reported an AUC of 0.76 in the internal-external validation, assessed through 10-fold nested cross-validation. See [12] for further detail about internal and internal-external validation.

As previously mentioned, our aim was then to illustrate the application of the proposed meta-validation methodology. To this purpose, we performed two experiments, to illustrate both steps of the above mentioned methodology. First, we considered the first step of the methodology, by means of a simulation

Feature	Italy-1	Italy-2	Italy-3	Spain	Brazil-1	Brazil-2	Brazil-3	Ethiopia
Age (years)	66.35 ± 0.1	54.38 ± 0.2	60.53 ± 0.2	56.68 ± 0.3	47.01 ± 0.0	42.87 ± 0.0	54.40 ± 0.1	59.73 ± 0.1
HCT (%)	38.20 ± 0.0	37.77 ± 0.1	39.67 ± 0.1	40.19 ± 0.1	41.47 ± 0.0	41.18 ± 0.0	40.75 ± 0.0	41.01 ± 0.0
HGB (g/dL)	13.21 ± 0.0	12.86 ± 0.0	13.11 ± 0.0	13.40 ± 0.0	13.82 ± 0.0	14.11 ± 0.0	13.74 ± 0.0	13.81 ± 0.0
MCH (pg/Cell)	29.62 ± 0.0	30.41 ± 0.0	29.51 ± 0.0	29.39 ± 0.0	29.37 ± 0.0	29.59 ± 0.0	29.60 ± 0.0	29.88 ± 0.0
MCHC (g Hb/dL)	34.49 ± 0.0	33.98 ± 0.0	33.00 ± 0.0	33.32 ± 0.0	33.31 ± 0.0	34.25 ± 0.0	33.68 ± 0.0	33.48 ± 0.0
MCV (fL)	85.72 ± 0.0	89.44 ± 0.1	89.41 ± 0.1	88.08 ± 0.1	88.19 ± 0.0	86.39 ± 0.0	87.90 ± 0.0	88.68 ± 0.0
RBC (10^{12}/L)	4.49 ± 0.0	4.25 ± 0.0	4.46 ± 0.0	4.58 ± 0.0	4.72 ± 0.0	4.78 ± 0.0	4.65 ± 0.0	4.74 ± 0.0
WBC (10^9/L)	9.81 ± 0.0	8.31 ± 0.1	9.53 ± 0.1	9.43 ± 0.1	6.70 ± 0.0	7.66 ± 0.0	6.30 ± 0.0	9.66 ± 0.1
PLT1 (10^9/L)	220.23 ± 0.5	204.00 ± 0.9	218.00 ± 0.7	220.07 ± 1.2	246.96 ± 0.1	239.92 ± 0.1	199.81 ± 0.4	259.03 ± 0.5
NE (%)	75.03 ± 0.1	67.54 ± 0.1	72.48 ± 0.1	72.81 ± 0.2	56.79 ± 0.0	61.02 ± 0.0	65.78 ± 0.1	68.81 ± 0.1
LY (%)	16.56 ± 0.1	21.90 ± 0.1	18.30 ± 0.1	17.96 ± 0.2	31.10 ± 0.0	27.57 ± 0.0	23.14 ± 0.1	21.76 ± 0.1
MO (%)	7.17 ± 0.0	8.86 ± 0.0	8.13 ± 0.0	7.87 ± 0.1	9.29 ± 0.0	8.64 ± 0.0	9.64 ± 0.0	6.76 ± 0.0
EO (%)	0.74 ± 0.0	1.23 ± 0.0	0.60 ± 0.0	1.00 ± 0.0	2.30 ± 0.0	2.27 ± 0.0	1.05 ± 0.0	2.11 ± 0.0
BA (%)	0.18 ± 0.0	0.46 ± 0.0	0.32 ± 0.0	0.36 ± 0.0	0.52 ± 0.0	0.48 ± 0.0	0.30 ± 0.0	0.56 ± 0.0
NET (10^9/L)	7.47 ± 0.0	5.62 ± 0.0	6.76 ± 0.0	7.20 ± 0.1	3.92 ± 0.0	4.82 ± 0.0	4.35 ± 0.0	7.13 ± 0.0
LYT (10^9/L)	1.63 ± 0.0	1.84 ± 0.0	1.82 ± 0.1	1.43 ± 0.0	2.01 ± 0.0	2.00 ± 0.0	1.31 ± 0.0	1.33 ± 0.0
MOT (10^9/L)	0.64 ± 0.0	0.73 ± 0.0	0.64 ± 0.0	0.69 ± 0.0	0.59 ± 0.0	0.63 ± 0.0	0.56 ± 0.0	0.52 ± 0.0
EOT (10^9/L)	0.06 ± 0.0	0.09 ± 0.0	0.05 ± 0.0	0.09 ± 0.0	0.15 ± 0.0	0.17 ± 0.0	0.06 ± 0.0	0.14 ± 0.0
BAT (10^9/L)	0.02 ± 0.0	0.03 ± 0.0	0.02 ± 0.0	0.03 ± 0.0	0.03 ± 0.0	0.03 ± 0.0	0.02 ± 0.0	0.04 ± 0.0
Sex (M/F)	65%/35%	68%/32%	63%/37%	53%/47%	43%/57%	53%47%	64%/36%	57%/43%
Suspect (Y/N)	48%/0%	42%/5%	82%/18%	85%/15%	0%/0%	0%/0%	0%/0%	50%/50%

Table 2: The quantitative features (along with units of measure, in parentheses) of the considered datasets: for each feature and dataset we report the mean and the 95% confidence interval. In regard to the Suspect feature, when the Yes (Y) and No (N) values do not sum up to 100%, the remaining percentage of values was missing.

experiment in which we generated from the original training dataset 100 random and 2 non-random (namely, using the data collected at the HSR during March/April as training set, and the data collected at the HSR during May, as well as the data collected at the IOG, as test sets) hold-out splits. Second, we considered the second step of the methodology, by means of the evaluation of the previously described ML model on the 8 external validation datasets. To illustrate the broad applicability of the proposed methodology, we also evaluated the correlation between dataset similarity and predictive performance: Statistical significance was evaluated using a confidence threshold of 95% ($\alpha = .05$).

3.3. Results

According to the first step of the proposed methodology, we first evaluated the correlation between similarity (measured through the Ψ metric) and the balanced accuracy of the ML model, on the basis of the 100 random splits and 2 non-random partitions (the IOG and HSR May partitions) of the original training set. The results are reported in the *potential robustness* diagram, shown in Figure 7. The top half of this diagram reports a scatter-plot of the 100 random

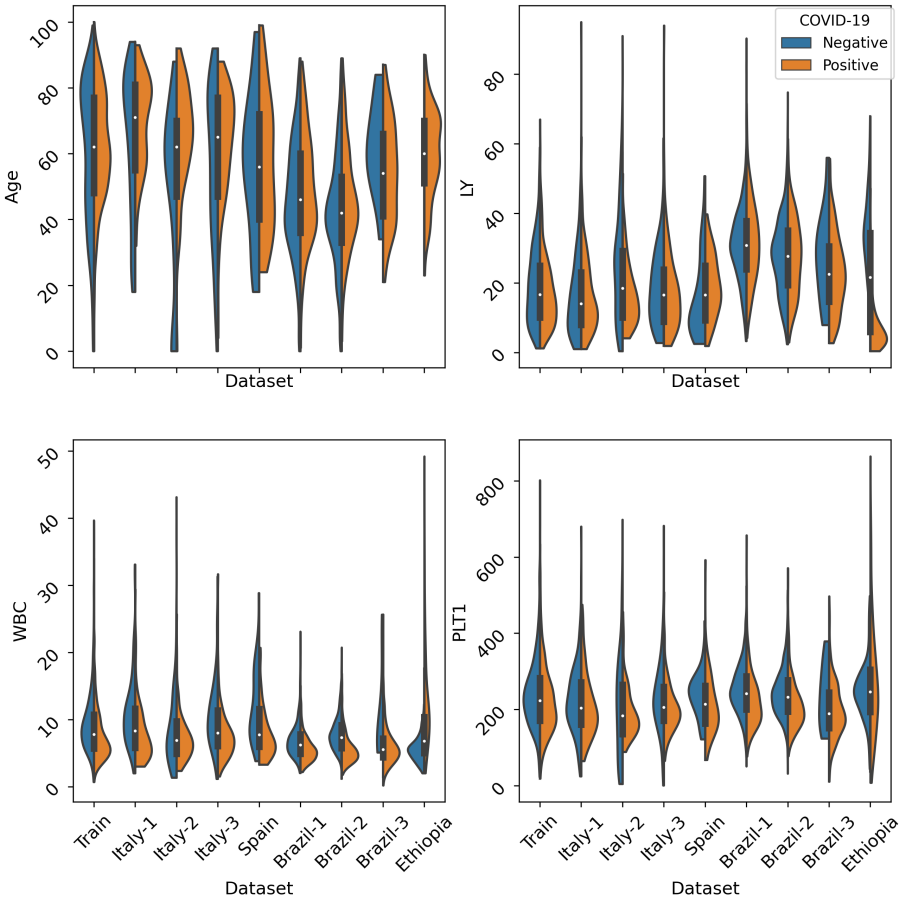

Figure 5: Violinplots of the most predictive features for the training set (train) and each of the external validation sets considered in this study. In clockwise order: age (years), lymphocyte count (%), white blood count (10^9/L), and platelet count (10^9/L).

splits: the average similarity was $.43 \pm .06$, while the average balanced accuracy was $.76 \pm .004$. In particular, the IOG partition was found to be very different from the rest of the training set, with a similarity of only .1 and an accuracy of .70, while the data collected at the HSR during May 2020 had a similarity of .43 (with respect to the data collected between March and April 2020) and an accuracy of .82. The bottom half of Figure 7 reports the regression line within the diagram proposed in Section 2. The dataset similarity and balanced accuracy were moderately correlated (Pearson $\rho = .38$) and the correlation was statistically significant ($p < .001$). The corresponding regression model had an angular coefficient of $b = 0.03$ and an intercept of $a = 0.76$, with $R^2 = 0.14$.

In regard to the second step of the methodology, as described in the previous section, we evaluated the ML model on 8 external validation datasets. The

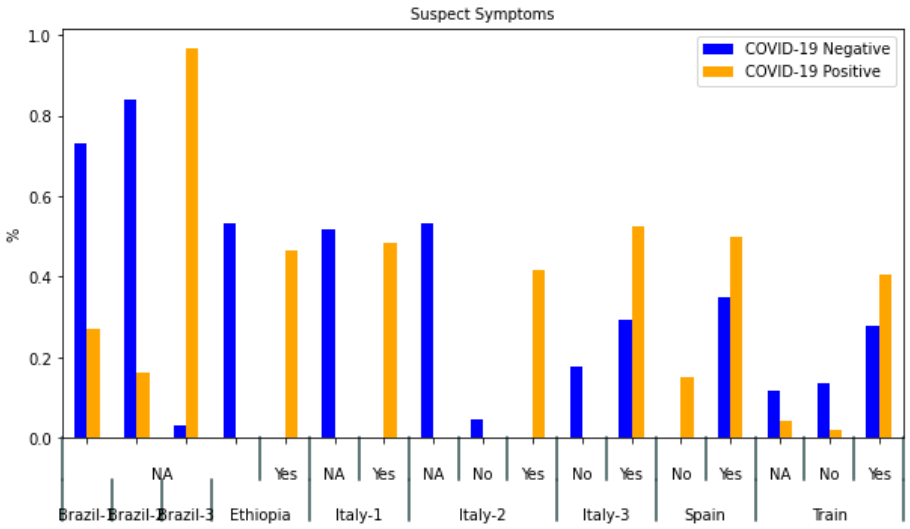

Figure 6: Prevalence of suspect symptoms, stratified by target and dataset.

Dataset	Type	Instances	CBC Analyzer	Missing rate (CBC + Age)	Missing Rate (Suspect)	Distance from training setting (km)	Collection Period
IOG	Internal	58	iSysmex XN-2000	1%	0%	7.5	February-April 2020
HSR-May	Internal	235	Sysmex XE 2100	28%	0%	0	May 2020
Italy-1	External	337	Sysmex XN-9000	0%	57%	15	March-April 2020
Italy-2	External	249	Sysmex XN-9000	0%	53%	36	March-April 2020
Italy-3	External	224	Sysmex XE 2100	3%	0%	0	November 2020
Spain	External	120	Roche XN 1000	0%	0%	1225	October 2020
Brazil-1	External	1301	NA	0%	100%	9500	February-June 2020
Brazil-2	External	2335	NA	0%	100%	9600	February-June 2020
Brazil-3	External	345	NA	0%	100%	9550	February-June 2020
Ethiopia	External	400	Beckman Coulter DXH	3%	0%	4930	January-March 2021

Table 3: Characteristics of the eight external validation datasets and two relevant partitions of the training set: the IOG dataset and the HSR-May dataset. In the Ethiopia dataset, the Age feature was missing for all the COVID-19 negative patients; while for the Italy-3 dataset, the formula features (i.e. NET, LYT, BAT, EOT, MOT and the respective percentage features) were missing for 11 patients.

performance of the ML model on the external validation datasets is reported in Table 4. Although the specificity was very high across all datasets (average $90\% \pm 6\%$), the model reported varying sensitivity (average $60\% \pm 19\%$), with higher performance on the Italian datasets (average F_2 score $87\% \pm 3\%$), and the worst performance obtained on the Brazilian datasets (average F_2 score $37\% \pm 4\%$); the decrease in performance was largely due to a large presence of false negative classifications: indeed, the model reported much lower sensitivity on the Brazilian datasets.

Table 4 also reports the values of the dataset similarity Ψ. In regard to the second step, the *external performance diagram*, reporting the correlation between the Ψ and, respectively, the AUC, the Net Benefit and the Brier Score, of the ML model on the external validation sets, is shown in Figure 8. The

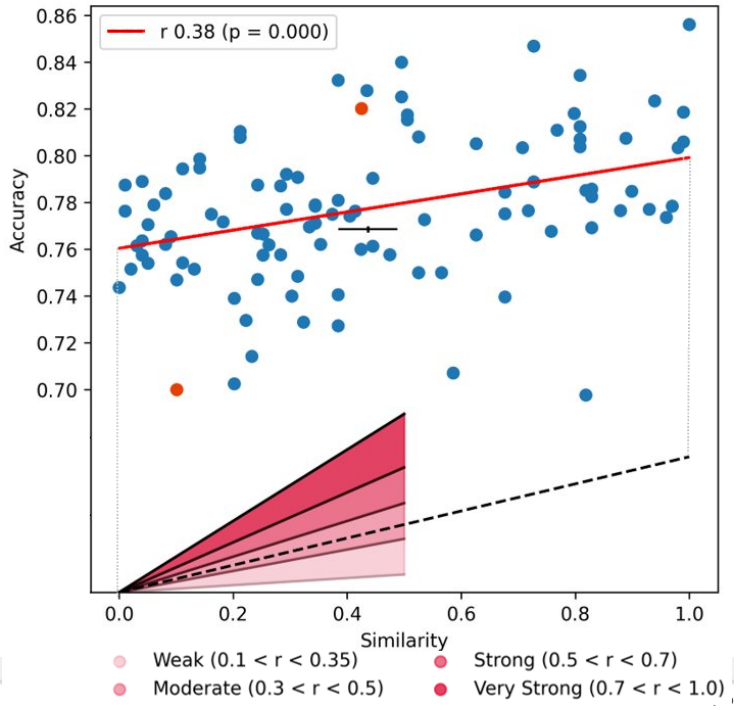

Figure 7: The *potential robustness diagram*, displaying the results of the "simulation of generalizability" step (i.e., step 1) for the study on the COVID-19 diagnosis. The correlation, shown by the dotted line, is moderate and statistically significant. Each circle represents a dataset from the repeated hold-out procedure; red circles represent two particular hold-out validation sets: the data collected at HSR in May 2020 (above) and the data collected at IOG (below). The black segment indicates the average similarity found among the partition datasets and its 95% confidence interval, indicating a 'moderate' similarity (see Table 1).

Dataset	Accuracy	Sensitivity	Specificity	AUC	Balanced Accuracy	F_2 Score	Brier Score	Standardized Net Benefit	Ψ
Italy-1	90%	91%	89%	97%	90%	91%	0.08	79%	0.439
Italy-2	93%	84%	99%	98%	91%	87%	0.08	83%	0.445
Italy-3	81%	85%	77%	89%	81%	83%	0.14	64%	0.447
Spain	68%	60%	81%	66%	71%	63%	0.27	50%	0.315
Brazil-1	77%	29%	95%	75%	62%	34%	0.16	16%	0.341
Brazil-2	86%	31%	97%	83%	64%	36%	0.10	15%	0.444
Brazil-3	39%	37%	91%	80%	64%	42%	0.39	37%	0.348
Ethiopia	79%	66%	90%	87%	78%	69%	0.15	56%	0.323

Table 4: The performance of the ML model on the external validation datasets.

correlation between the AUC and the dataset similarity was very strong ($r = .74$) and significant ($p = .035$); the correlation between the Net Benefit and dataset similarity was moderate ($r = .39$) but not significant ($p = 0.345$); while the correlation between the Brier score and dataset similarity was strong ($r = .66$) but not significant ($p = .076$), likely due to the relatively small number of

datasets. Thus, in light of the results of the first step of the methodology, we can see that the findings of the external validation confirm the observed moderate impact of data heterogeneity on model performance, since the best reported performance was mostly associated with the more similar external datasets (i.e. the Italian dataset).

Based on what said in Section 2, the model can be considered externally validated, as, for all three performance metrics, at least one external dataset was associated with slight similarity and acceptable (or better) performance. Furthermore, for all three performance metrics, most external validation datasets could be considered of sufficient cardinality: indeed, most datasets exceeded the MSS for the three performance metrics, whereas only the Spain dataset was associated with a dataset cardinality smaller than the MSS for all the three performance metrics. That said, the variability in the observed results prompts for further discussion in the next section.

4. Discussion

The external validation of ML models is increasingly being proposed as the main (and only) means to certify the supposed validity of the model on (virtually any) unseen data [19, 59]. However, as also the findings shown above illustrate, the result from an external validation cannot guarantee reliability *per se* [25]: if the external validation had been performed only on the Italy-2 dataset, where the diagnostic model performed even better than on the original test set and exhibited very high AUC scores, the external validation would have been considered a clear success; if, conversely, the external validation had been performed on the Spain dataset, where the model accuracy dropped by losing almost 30 percent points (see Figure 8) such a claim would have seemed inflated at best. Thus, as the old saying runs, only time can tell and, as we also argued elsewhere [14], only "eating the pudding can prove its quality". However, we proposed to take some *informed* guess by performing a meta-validation procedure, including the external validation itself.

Indeed, when considering external validation, two main questions must be addressed: "is my validation *actually* reliable enough?"; and "Is my model *actually* valid?". The first question is addressed by the first step of the procedure we propose (see Section 2), by evaluating the extent the validation data are different from training data *and* the extent the model performance seems to be susceptible in that respect. The *potential robustness diagram* is a visual aid to address these latter aspects. On the other hand, the second question is addressed by the second step of the procedure we propose (see Section 2). To this respect, the vertical axis of the *external performance diagram* helps understand how much the validation is *actually external* (hence, how conservative), while the horizontal axis of this diagram is suggestive of the comprehensive validity of the model on the external dataset(s).

Lastly, since accuracy scores are always probabilistic estimates (despite the fact they are seldom presented as such, e.g., with confidence intervals), the

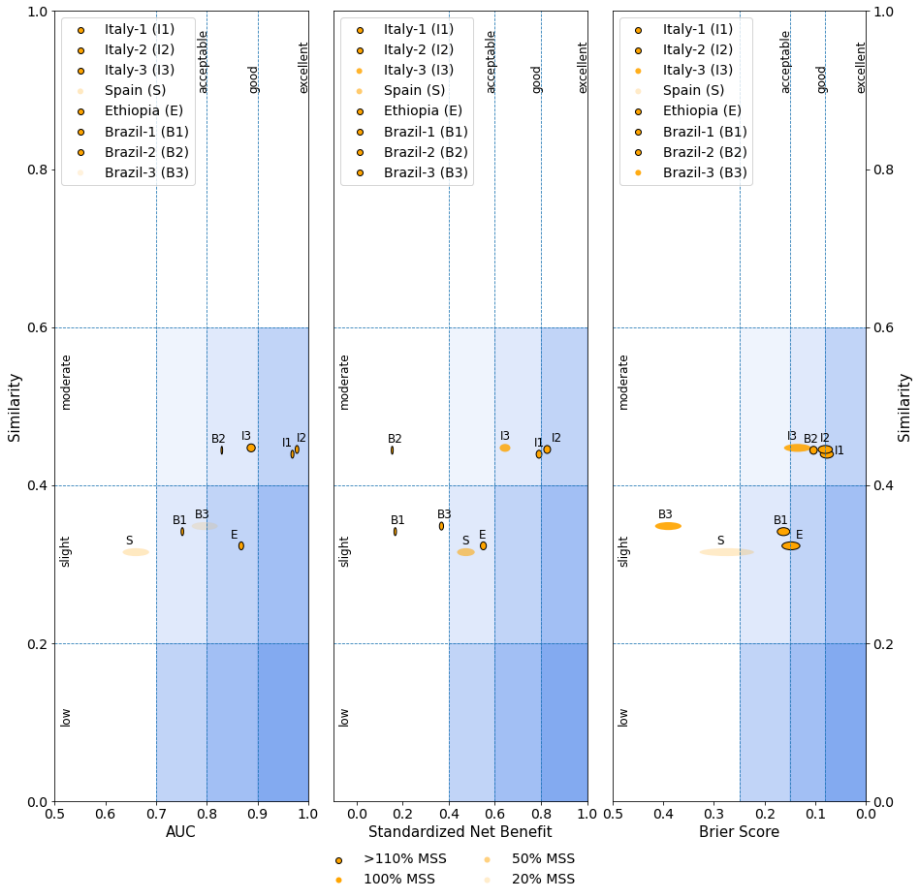

Figure 8: The *external performance diagram*, displaying the results from the external validation study on the COVID-19 diagnosis (second step of the meta-validation procedure). Information about the MSS is rendered in terms of hue brightness. The width of the ellipses is equal to the width of the 95% confidence interval w.r.t. the given performance metrics.

external performance diagram gives also clues about the *significance* of the performance scores, in terms of the degree to which the 'minimum sample size' requirement has been met (or not).

We make the point that combining information about similarity and performance susceptibility together (first step), and about similarity and statistical significance (second step, and related diagrams) is important to allow the qualitative interpretation of the *representativeness* of the external datasets and their role to make claims of robustness reliable.

As we have seen in the COVID-19 case study, many factors can influence the representativeness of datasets for assessing the robustness of a model in a given classification task: differences in testing equipment [43, 24] (cf. the concepts of *harmonization* and *analytical variability*), in reference ranges [15, 40] (cf.

ethnic variability), in disease manifestations (cf. phenotypic variability) and how humans react to diseases, also due to contextual factors (cf., *biological intra-individual variability* and the *clinical inter-group variability*) [21] make the reference population extremely vast and various, from which also very dissimilar datasets can be drawn to challenge the model's performance.

In this respect, our work is among the first ones to assert the obvious, but seriously neglected fact, that *external datasets are not all the same* [50]. Thus, given a classification task, researchers who are diligent in the external validation of their models should acquire multiple datasets and test the performance of their models over these datasets. Furthermore, in this paper we make the point of the need to also assess how *diverse* the external datasets are with respect to the training set [7, 13, 53] , and how *large* the external datasets are with respect to what would be required to achieve reliable performance estimates [49, 55, 64].

For instance, with reference to our case study and, in particular, to Figure 8, we can see that the validated model performed well (either 'good' or 'excellent') on three/four datasets that were 'moderately' dissimilar and of adequate sample size (as indicated by the border of their corresponding circles). As a *general rule of thumb*, an *external validation can be considered successful* when the model exhibits an *at-least-good* performance on (the majority of) the *at-least-moderately-similar* external dataset(s) at hand (with respect to the training set). However, validating the model on multiple external datasets allow for some more fine-grained considerations. In our COVID-19 case, for instance, the performance exhibited on the Ethiopian dataset (E in Figure 8) is more interesting than it might seem at first sight, because it was obtained on one of the most diverse datasets, even more different than the Brazilian ones (B1, B2 and B3 in Figure 8), which, on the other hand, are the most distant ones (geographically speaking). Conversely, the model performance on the Brazilian datasets suggests great caution in adopting the validated model on any *comparable* datasets that: *(i)* exhibit the same imbalance or similar data distributions (in particular the patients in the Brazil-1 and Brazil-2 datasets were of significantly younger age); *(ii)* are collected from different or unknown equipment; *(iii)* or similarly lack important predictive features (in the case of the Brazilian cohorts, the *Suspect* feature, recognized as one of the most important predictive features [42], was systematically missing). With respect to this latter point, and looking at Figure 6, we can also observe that the ML model exhibited better performance on those datasets where the diagnosis agreed with the reported symptoms. In light of this latter observation, the performance obtained on the Spain and Italy-3 datasets can be considered interesting, as both datasets contained a significant portion of instances whose symptoms could be considered misleading with respect to the assigned diagnosis.

The COVID-19 case study presented here represents a real-world case, in which we tested a state-of-the-art model on a delicate and relevant task (like COVID-19 early diagnosis) with data coming from settings thousands of kilometres apart in three continents. The application of the proposed methodology allowed us to externally validate the above model, as this latter achieves adequate performance levels in all the three dimensions considered (see Figure 8).

These results should be interpreted also in light of the results of the first step of the methodology. Since susceptibility to data heterogeneity was found to be moderate (see Figure 7), caution should be exercised when the model is to be applied to other external (and different) datasets. We made the point (and proved it) that these aspects play an important role in the sound evaluation of medical ML [11].

This case study allows us to recognize the main advantages and limits of the meta-validation method proposed in this contribution. The main advantage regards the simple and lean nature of the two-step method, which yet does not sacrifice comprehensiveness, as it allows to evaluate a classification model in terms of complementary dimensions, like discrimination, calibration and utility. Its qualitative approach to the interpretation of performance scores is both and advantage and a limit: a model can be very good in discriminative terms while being poor in terms of utility; it is then up to the designer, or the prospective user, to decide what dimensions are more important according to their purposes and needs. Our method, particularly so through the first step of the methodology, also allows to get a qualitative idea of the impact of dataset heterogenity on the model performance and robustness: while no one can exactly predict how a model will perform with some unseen data, our *potential robustness* diagram allows to get an idea of the susceptibility of the model to data heterogeneity, in terms of data similarity, and this is the first time such an approach is pursued.

In regard to possible future works, first of all, we note that while in the first step of the proposed methodology we focused on *linear* relationships between the dataset similarity and model performance, the proposed approach could in principle be extended also to more general, non-linear relationships. Obviously, such generalization would necessitate the development of different visualizations for the *potential robustness* diagram. Similarly, the second step of the proposed methodology could in principle be adapted to other performance metrics. A possible research direction, then, would regard the development of novel MSS formulas and the definition of appropriate performance thresholds.

Finally, we mention that even though here we focused on the validation of ML models, assessing similarity and cardinality of an external dataset can also be useful for other phases that are not directly related to validation. For instance, a similarity score could, in principle, be used to efficiently improve the model over time and cope with changes in the phenomenon to be classified [48] (e.g., disease manifestation, for changes in the catchment area, case mix, data collection policies, employed tests or the mutating pathogen): to this aim, efficiency would be achieved by using similarity scoring to identify the most dissimilar data points that the model should learn to classify, and the by focusing on those data. Also procurement managers can leverage similarity scores to inform their decisions: to this aim, they could ask vendors to compare their training set (not necessarily disclosed) with a representative sample of local instances, so as to assess the similarity between these two datasets in terms of some metric (e.g., in case of tabular data, the Data Representativeness Criterion [53] or the *Degree of Correspondance* [13] adopted in this article). If the datasets are found to be too *dissimilar*, any statement on the accuracy of

the model should be taken with extreme caution and require further inquiries.

5. Conclusions and final recommendations

The external validation of a medical machine learning model is very important for a number of reasons [19, 59]. This is also the case because an external validation corroborates the reputation on the model, and hence the users' trust and performance expectancy, which are known to be positively correlated with the behavioral intention to adopt and use the system in the actual practice [22, 34]. However, for external validation to provide a sounder basis for more reliable estimates (than internal validation) of the prospective performance of the model on new, unseen cases in multiple setting, just to make it "external", that is based on data coming from other settings than the training one, is simply not enough.

For this purpose, in this paper, we proposed a meta-validation methodology to assess any validation procedure through two qualitative, graphical tools that allow to evaluate the robustness w.r.t. data similarity (i.e., the *potential robustness diagram*) as well as the results of an external validation, in light of data similarity and data cardinality (i.e., the *external performance diagram*). Furthermore, through a real-world case study, we described the application of our methodology and have shown that the performance of an accurate model to detect COVID-19 from routine blood tests (whose analytical variability is negligible across laboratories from all over the world [65]) significantly degrades when these tests are taken in different settings, by means of different equipment, or on heterogeneous populations. These results corroborate and extend the results obtained in the previous literature [53, 55] studying the impact of data similarity and data cardinality on model generalization. Furthermore, we emphasize that the model we "validated" in this article, grounds on the most stable and less variable hematochemical exam [65], the complete blood count: and yet, the correlation between accuracy and similarity that we reported should serve as a warning sign that reproducing good performances across very heterogeneous settings can be overambitious and unrealistic. The same reasoning obviously applies, and if possible even exacerbated, in the case of models that rely on less stable tests, such as imaging [50].

In light of this study we can finally share some recommendations when developing a ML model in medical settings (and other similarly critical domains):

1. Do an external validation;
2. If doing an external validation is not feasible, perform a hold-out internal validation where the hold-out dataset has been chosen among a not-so-small set of candidates so as to be the most diverse one, by computing some apt similarity score (e.g., the *Degree of Correspondance* Ψ adopted herein; in this case the similarity should be low or extremely low – see Table 1). Alternatively, perform a cross-validation but report the average performance on the most diverse 20% of the test partitions. As shown in Figure 7 (see its leftmost part), doing so does not prevent "good"

scores, but it may yield more conservative (and hence reliable) estimates of model's robustness;

3. When doing the external validation, take into account the similarity between the training set and the external validation set, and the cardinality of the latter. Be aware that if similarity is substantial or higher (see Table 1), or if the cardinality is less than the minimum sample size, then the external validation may not yield estimates of future performance of sufficient reliability; Visual aids like the *potential robustness diagram* and the *external performance diagram* depicted in Figure 7 and Figure 8, respectively, can be conveniently used to support result interpretation: the code to generate them is available online[3] and free to use for any interested researcher.
4. And then, again: do an external validation, even if you won't like the results.

Acronyms and abbreviations

- BA: Basophils
- CBC: Complete Blood Count
- DAC: Data Agreement Criterion
- DRC: Data Representativeness Criterion
- ED: Emergency Department
- EO: Eosinophils
- HCT: Hematocrit
- HGB: Hemoglobin
- HSR: Hospital San Raffaele
- IOG: Istituto Ortopedico Galeazzi
- LY: Lymphocytes
- MCH: Mean Corpuscular Hemoglobin
- MCHC: Mean Corpuscular Hemoglobin Concentration
- MCV: Mean Corpuscular Volume
- ML: Machine Learning
- NE: Neutrophils

[3] `https://github.com/AndreaCampagner/qualiMLpy`

- PLT1: Platelets
- RBC: Red Blood Cells
- RBF: Radial Basis Function
- WBC: White Blood Cells

Acknowledgments and Declarations

Ethics Approval:. Research involving human subjects complied with all relevant national and international regulations, institutional policies and is in accordance with the tenets of the Helsinki Declaration (as revised in 2013), and was approved by the authors' Institutional Review Board (70/INT/2020).

Funding:. Funding sources had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding authors had full access to all data in the study and had final responsibility for the decision to submit for publication.

Competing Interests:. The authors report no competing interest.

References

[1] Kartik Ahuja. Estimating kullback-leibler divergence using kernel machines. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 690–696. IEEE, 2019.

[2] Lucinda Archer, Kym IE Snell, Joie Ensor, Mohammed T Hudda, Gary S Collins, and Richard D Riley. Minimum sample size for external validation of a clinical prediction model with a continuous outcome. *Statistics in Medicine*, 40(1):133–146, 2021.

[3] Ali Abbasian Ardakani, Alireza Rajabzadeh Kanafi, U Rajendra Acharya, Nazanin Khadem, and Afshin Mohammadi. Application of deep learning technique to manage covid-19 in routine clinical practice using ct images: Results of 10 convolutional neural networks. *Computers in Biology and Medicine*, 121:103795, 2020.

[4] Andrew L Beam, Arjun K Manrai, and Marzyeh Ghassemi. Challenges to the reproducibility of machine learning models in health care. *Jama*, 323(4):305–306, 2020.

[5] SE Bleeker, HA Moll, EW Steyerberg, ART Donders, Gerarda Derksen-Lubsen, DE Grobbee, and KGM Moons. External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology*, 56(9):826–832, 2003.

[6] Sylvain Boltz, Eric Debreuve, and Michel Barlaud. knn-based high-dimensional kullback-leibler distance for tracking. In *Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'07)*, pages 16–16. IEEE, 2007.

[7] Nicolas Bousquet. Diagnostics of prior-data agreement in applied bayesian analysis. *Journal of Applied Statistics*, 35(9):1011–1029, 2008.

[8] A Allen Bradley, Stuart S Schwartz, and Tempei Hashino. Sampling uncertainty and confidence intervals for the brier score and brier skill score. *Weather and Forecasting*, 23(5):992–1006, 2008.

[9] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of ICPR 2010*, pages 3121–3124. IEEE, 2010.

[10] Luca Brunese, Francesco Mercaldo, Alfonso Reginelli, and Antonella Santone. Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. *Computer Methods and Programs in Biomedicine*, 196:105608, 2020.

[11] Federico Cabitza and Andrea Campagner. The need to separate the wheat from the chaff in medical informatics. *International Journal of Medical Informatics*, 2021.

[12] Federico Cabitza, Andrea Campagner, Davide Ferrari, Chiara Di Resta, Daniele Ceriotti, Eleonora Sabetta, Alessandra Colombini, Elena De Vecchi, Giuseppe Banfi, Massimo Locatelli, et al. Development, evaluation, and validation of machine learning models for covid-19 detection based on routine blood tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 59(2), 2021.

[13] Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. As if sand were stone. new concepts and metrics to probe the ground on which to build trustable ai. *BMC Medical Informatics and Decision Making*, 20(1):1–21, 2020.

[14] Federico Cabitza and Jean-David Zeitoun. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Annals of translational medicine*, 7(8), 2019.

[15] Anna Carobene, Annette Gathoni Kiarie, Mizan Tsegaye, Michela Seghezzi, Sabrina Buoro, and Ritah Wanja Mburugu. A very uncommon haemoglobin value resulting from a severe acute malnutrition in a 16-month-old child in ethiopia. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 59(3):20200364, 2020.

[16] Gavin C Cawley and Nicola LC Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, 11:2079–2107, 2010.

[17] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14(1):1–22, 2021.

[18] Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[19] Gary S Collins, Joris A de Groot, Susan Dutton, Omar Omar, Milensu Shanyinde, Abdelouahid Tajar, Merryn Voysey, Rose Wharton, Ly-Mee Yu, Karel G Moons, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology*, 14(1):1–11, 2014.

[20] Gary S Collins, Emmanuel O Ogundimu, and Douglas G Altman. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in medicine*, 35(2):214–226, 2016.

[21] Abdurrahman Coskun, Federica Braga, Anna Carobene, Xavier Tejedor Ganduxe, Aasne K Aarsand, Pilar Fernández-Calle, Jorge Díaz-Garzón Marco, William Bartlett, Niels Jonker, Berna Aslan, et al. Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 58(1):25–32, 2020.

[22] José Manuel Ortega Egea and María Victoria Román González. Explaining physicians' acceptance of ehcr systems: An extension of tam with trust and risk factors. *Computers in Human Behavior*, 27(1):319–332, 2011.

[23] Davide Ferrari, Andrea Motta, Marta Strollo, Giuseppe Banfi, and Massimo Locatelli. Routine blood tests as a potential diagnostic tool for COVID-19. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 58(7), 2020.

[24] Carlo Franzini. Relevance of analytical and biological variations to quality and interpretation of test results: examples of application. *Ann. Ist. Super. Sanità*, 31(1):9–13, 1995.

[25] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health*, 2(9):e489–e492, 2020.

[26] Vicente García, Ramon A Mollineda, and J Salvador Sánchez. Theoretical analysis of a performance measure for imbalanced data. In *2010 20th International Conference on Pattern Recognition*, pages 617–620. IEEE, 2010.

[27] Vicente García, José Salvador Sánchez, and Ramón Alberto Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2012.

[28] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19:513–520, 2006.

[29] Tinotenda A Harahwa, Thomas Ho Lai Yau, Mae-Sing Lim-Cooke, Salah Al-Haddi, Mohamed Zeinah, and Amer Harky. The optimal diagnostic methods for covid-19. *Diagnosis*, 7(4):349–356, 2020.

[30] Tina Hernandez-Boussard, Selen Bozkurt, John PA Ioannidis, and Nigam H Shah. Minimar (minimum information for medical ai reporting): developing reporting standards for artificial intelligence in health care. *Journal of the American Medical Informatics Association*, 27(12):2011–2015, 2020.

[31] Hamish Huggard, Yun Sing Koh, Gillian Dobbie, and Edmond Zhang. Detecting concept drift in medical triage. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1733–1736, 2020.

[32] Asif Iqbal Khan, Junaid Latief Shah, and Mohammad Mudasir Bhat. Coronet: A deep neural network for detection and diagnosis of covid-19 from chest x-ray images. *Computer Methods and Programs in Biomedicine*, 196:105581, 2020.

[33] Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational statistics & data analysis*, 53(11):3735–3745, 2009.

[34] Seok Kim, Kee-Hyuck Lee, Hee Hwang, and Sooyoung Yoo. Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (emr) using the unified theory of acceptance and use of technology (utaut) in a tertiary hospital. *BMC medical informatics and decision making*, 16(1):1–12, 2015.

[35] Inke R König, JD Malley, C Weimar, H-C Diener, and A Ziegler. Practical experiences on the necessity of external validation. *Statistics in medicine*, 26(30):5499–5511, 2007.

[36] Wouter M Kouw, Marco Loog, Lambertus W Bartels, and Adriënne M Mendrik. Learning an mr acquisition-invariant representation using siamese neural networks. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 364–367. IEEE, 2019.

[37] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[38] Jake Lever, Martin Krzywinski, and Nicole Altman. Model selection and overfitting. *Nature Methods*, 13:703–704, 2016.

[39] Xiangchun Li, Sheng Zhang, Qiang Zhang, Xi Wei, Yi Pan, Jing Zhao, Xiaojie Xin, Chunxin Qin, Xiaoqing Wang, Jianxin Li, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *The Lancet Oncology*, 20(2):193–201, 2019.

[40] E-M Lim, George Cembrowski, M Cembrowski, and G Clarke. Race-specific wbc and neutrophil count reference intervals. *International journal of laboratory hematology*, 32(6p2):590–597, 2010.

[41] Jayawant N Mandrekar. Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, 5(9):1315–1316, 2010.

[42] Cristina Menni, Ana Valdes, Maxim B Freydin, Sajaysurya Ganesh, Julia El-Sayed Moustafa, Alessia Visconti, Pirro Hysi, Ruth CE Bowyer, Massimo Mangino, Mario Falchi, et al. Loss of smell and taste in combination with other symptoms is a strong predictor of covid-19 infection. *MedRxiv*, 2020.

[43] W Greg Miller. Harmonization: its time has come. *Clinical Chemistry*, 63(7), 2017.

[44] Tulin Ozturk, Muhammed Talo, Eylul Azra Yildirim, Ulas Baran Baloglu, Ozal Yildirim, and U Rajendra Acharya. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine*, 121:103792, 2020.

[45] Menelaos Pavlou, Chen Qu, Rumana Z Omar, Shaun R Seaman, Ewout W Steyerberg, Ian R White, and Gareth Ambler. Estimation of required sample size for external validation of risk models for binary outcomes. *Statistical Methods in Medical Research*, page 09622802211007522, 2021.

[46] Rodolfo M Pereira, Diego Bertolini, Lucas O Teixeira, Carlos N Silla Jr, and Yandre MG Costa. Covid-19 identification in chest x-ray images on flat and hierarchical classification scenarios. *Computer Methods and Programs in Biomedicine*, 194:105532, 2020.

[47] Timothy B Plante, Aaron M Blau, Adrian N Berg, Aaron S Weinberg, Ik C Jun, Victor F Tapson, Tanya S Kanigan, and Artur 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. *Journal of medical Internet research*, 22(12):e24048, 2020.

[48] Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younes Bennani. *Advances in domain adaptation theory*. Elsevier, 2019.

[49] Richard D Riley, Thomas PA Debray, Gary S Collins, Lucinda Archer, Joie Ensor, Maarten van Smeden, and Kym IE Snell. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Statistics in Medicine*, 2021.

[50] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[51] Sherri Rose. Machine learning for prediction in electronic health data. *JAMA network open*, 1(4):e181404–e181404, 2018.

[52] Valentin Rousson and Thomas Zumbrunn. Decision curve analysis revisited: overall net benefit, relationships to roc curve analysis, and application to case-control studies. *BMC medical informatics and decision making*, 11(1):1–9, 2011.

[53] Evelien Schat, Rens van de Schoot, Wouter M Kouw, Duco Veen, and Adriënne M Mendrik. The data representativeness criterion: Predicting the performance of supervised classification based on data set similarity. *Plos one*, 15(8):e0237009, 2020.

[54] Ian Scott, Stacy Carter, and Enrico Coiera. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health & Care Informatics*, 28(1), 2021.

[55] Kym IE Snell, Lucinda Archer, Joie Ensor, Laura J Bonnett, Thomas PA Debray, Bob Phillips, Gary S Collins, and Richard D Riley. External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *Journal of clinical epidemiology*, 135:79–89, 2021.

[56] Andrew AS Soltan, Samaneh Kouchaki, Tingting Zhu, Dani Kiyasseh, Thomas Taylor, Zaamin B Hussain, Tim Peto, Andrew J Brent, David W Eyre, and David 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. *The Lancet Digital Health*, 2020.

[57] Xing Song, SL Alan, John A Kellum, Lemuel R Waitman, Michael E Matheny, Steven Q Simpson, Yong Hu, and Mei Liu. Cross-site transportability of an explainable artificial intelligence model for acute kidney injury prediction. *Nature communications*, 11(1):1–12, 2020.

[58] Ewout W Steyerberg, Sacha E Bleeker, Henriëtte A Moll, Diederick E Grobbee, and Karel GM Moons. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of clinical epidemiology*, 56(5):441–447, 2003.

[59] Ewout W Steyerberg and Frank E Harrell Jr. Prediction models need appropriate internal, internal-external, and external validation. *Journal of clinical epidemiology*, 69:245, 2016.

[60] Ewout W Steyerberg, Frank E Harrell Jr, Gerard JJM Borsboom, MJC Eijkemans, Yvonne Vergouwe, and J Dik F Habbema. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, 54(8):774–781, 2001.

[61] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

[62] Ben Van Calster, Daan Nieboer, Yvonne Vergouwe, Bavo De Cock, Michael J Pencina, and Ewout W Steyerberg. A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of clinical epidemiology*, 74:167–176, 2016.

[63] Duco Veen, Diederick Stoel, Naomi Schalken, Kees Mulder, and Rens Van de Schoot. Using the data agreement criterion to rank experts' beliefs. *Entropy*, 20(8):592, 2018.

[64] Yvonne Vergouwe, Ewout W Steyerberg, Marinus JC Eijkemans, and J Dik F Habbema. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *Journal of clinical epidemiology*, 58(5):475–483, 2005.

[65] Matteo Vidali, Anna Carobene, Sara Apassiti Esposito, Gavino Napolitano, Alessandra Caracciolo, Michela Seghezzi, Giulia Previtali, Giuseppe Lippi, and Sabrina Buoro. Standardization and harmonization in hematology: Instrument alignment, quality control materials, and commutability issue. *International Journal of Laboratory Hematology*, 2020.

[66] G. Wang, X. Liu, and J. Shen. 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*, 2021.

[67] Jiangpeng Wu, Pengyi Zhang, Liting Zhang, Wenbo Meng, Junfeng Li, Chongxiang Tong, Yonghong Li, Jing Cai, Zengwei Yang, Jinhong Zhu, et al. Rapid and accurate identification of covid-19 infection through machine learning based on clinical available blood test results. *medRxiv*, 2020.

[68] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten, Darren L Dahly, Johanna AA

Damen, Thomas PA Debray, et al. Prediction models for diagnosis and prognosis of COVID-19: systematic review and critical appraisal. *bmj*, 369, 2020.

[69] He S Yang, Yu Hou, Ljiljana V Vasovic, Peter AD Steel, Amy Chadburn, Sabrina E Racine-Brzostek, Priya Velu, Melissa M Cushing, Massimo Loda, Rainu Kaushal, et al. Routine laboratory blood tests predict sars-cov-2 infection using machine learning. *Clinical chemistry*, 66(11):1396–1404, 2020.

[70] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 2020.