Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240535

How Trueness of Clinical Decision Support Systems Based on Machine Learning Is Assessed?

Alex POIRON [a], Sandie CABON [a,1] and Marc CUGGIA[a]
[a] *Univ Rennes, CHU Rennes, INSERM, LTSI-UMR 1099, F-35000, Rennes, France*

Abstract. The application of machine learning algorithms in clinical decision support systems (CDSS) holds great promise for advancing patient care, yet practical implementation faces significant evaluation challenges. Through a scoping review, we investigate the common definitions of ground truth to collect clinically relevant reference values, as well as the typical metrics and combinations employed for assessing trueness. Our analysis reveals that ground truth definition is mostly not in accordance with the standard ISO expectation and that used combination of metrics does not usually cover all aspects of CDSS trueness, particularly neglecting the negative class perspective.

Keywords. clinical decision support system, machine learning, evaluation, trueness, scoping review

1. Introduction

Interest in applying machine learning algorithms in clinical settings, especially through Clinical Decision Support Systems (CDSS), is on the rise, with aims to enhance patient care. However, despite this interest, the adoption of such algorithms for patient use remains limited in practice. This is partly due to the fact that there is a mismatch between the evaluation required to obtain certification for real clinical integration, and the evaluation carried out at the proof-of-concept stage. The transition between these two stages requires additional development creating a bottleneck effect [1]. To address this, there is a need to identify objective criteria to evaluate the readiness of these solutions for clinical integration. We know that this assessment is multi-factorial and complex [2]. By studying recent regulations and guidelines, we have identified an initial list of important criteria. In this article, we focus on the criterion of trueness, as the evaluation of agreement between predicted and reference values of a CDSS. It is of great importance in the context of learning-based methods, as it is the one that is optimized when training and evaluating an AI-based CDSS. By the mean of a scoping review, we explore how ground truth is commonly defined to gather clinically relevant reference values and what metrics and combinations are usually used to evaluate trueness. We link these practices with what is required for certification and identify potential areas for improvement.

[1] Corresponding Author: Sandie CABON; E-mail: sandie.cabon@univ-rennes.fr.

2. Methods

2.1. Criteria identification

There is currently no standardized set of criteria for evaluating a learning-based CDSS. To conduct this review, we synthesized a list of main criteria from four sources. First, we examined the European Artificial Intelligence (AI) Act [3]. Although not health-specific, it categorizes CDSS as "high-risk systems" and stipulates requirements in Chapter 2. We also extracted requirements from Annexes XIV and XV of the European Medical Devices Regulation (MDR)[4], focusing on clinical evaluation and investigation. We also considered FUTURE-AI which proposed six guiding principles for trustworthy AI in medicine [5]. Lastly, we analyzed the World Health Organization's (WHO) guidelines on AI ethics and governance for health [6] and extracted keywords from Chapters 4 to 9. Ultimately, we organized these requirements into ten main criteria: Trueness, Robustness, Data Quality, Transparency, Human Oversight, Ethics, Clinical Utility, Data Privacy and Cybersecurity, Lifecycle Monitoring, and Regulatory Compliance. A figure summarizing this process can be found at [7].

2.2. Focus on trueness

In this paper, a focus on "Trueness" is made. It relates to three collected requirements: "The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared" [3], "ensure that the technologies were accurate and effective." [6] and "establishing the safety and performance of the device" [4]. According to ISO 5725-1:2023, "accuracy" combines "Trueness" and "Precision". "Trueness" is defined as "the closeness of agreement between the expectation of test results and a true value," focusing on systemic errors. On the other hand, "Precision" (i.e., not the metric) refers to "the closeness of agreement between independent test results obtained under stipulated conditions" and "depends only on the distribution of random errors and does not relate to the true value". "Precision" will be addressed in the Robustness criterion and is not covered in this paper. As the true value is not commonly available, it is substituted by the "accepted reference value". Therefore, when evaluating Trueness, we can consider it from two angles: firstly, by examining the process of defining the "accepted reference value," and secondly, by gathering metrics use to measure systemic errors. With this definition, we examined how the criterion is portrayed in the literature, examining the ground truth's definition and relation to clinical practice. Then, we analyzed the combinations of metrics employed to evaluate closeness of the predicted values to it.

2.3. Studies selection process for scoping review

We curated our review from a set of articles sourced through the PubMed database using the keywords "clinical decision support system", "machine learning" and "deep learning." We focused on articles published in 2022 to capture recent advancements in the field, excluding reviews. Due to the large volume of articles (n=265), a random selection maintaining proportional representation of various types of health data (tabular data, images, signals, etc.) was conducted. Ultimately, 77 articles were kept for analysis. All studies except one feature CDSS at the pre-clinical testing stage of technological readiness and a broad spectrum of clinical applications is covered (e.g., stroke prediction, hospital readmission risk). Details are available at [7].

3. Results

3.1. Definition of the ground truth

According to ISO 5725-1:2023, an accepted reference value can be: a) a theoretical or established value, based on scientific principles; b) an assigned or certified value, based on experimental work of some national or international organization; c) a consensus or certified value, based on collaborative experimental work under the auspices of a scientific or engineering group; d) the expectation, i.e., the mean of a specified population of measurements when a), b), and c) are not available [8]. In our review, 42.9% of the studies can be classified as a) and 15.6% as c), when multiple experts annotated all the data and consensus was found to build the ground truth. No correspondence could be made for 41.5% of the studies. Among them, we found 1.3% where multiple experts annotate separately a part of the data and had a consensus on another part, 11.7% where multiple experts annotated independent data batches, 6.5% where annotations are made by a single expert, 2.6% where ground truth is obtained without clear connection to a clinical practice, and 19.5% where the process to define ground truth was not directly mentioned.

3.2. Metrics used and their combinations

Figure 1 summarizes all metrics and combinations encountered for binary classification, multi-classification and regression tasks. For three studies, no metrics were computed. Regarding binary classification in **1A**, as a first remark, a wide variety of combinations was retrieved (39 in total, combining one to nine metrics). 14.3% only used one summary metric. Most evaluations focus on the balance between negative and positive performances using summary metrics (e.g., AUC-ROC, Acc) and on performance for the positive class (e.g., Prec, Rec). From six combined metrics onwards (36.6% of studies), there is a growing interest in evaluating performance on the negative class as well. For multi-classification in **1B**, most metrics are used in a binary representation of the problem, employing either the “one-vs-one” (most cases) or “one-vs-all” strategy. Two studies propose overall evaluation through “macro” or “micro” computations. One of them also used weighted-macro for F1 and Prec, allowing for weighting based on the representation of each class. Concerning regression in **1C**, 54.6% of the studies use only one metric to evaluate trueness. In this specific case, the Mean Absolute Error (MAE) was the most frequently used. All these combinations are also detailed in Annex [7].

4. Discussion

Regarding ground truth, a significant portion was defined either based on documented clinical practice or using consensual value obtained from multiple experts. However, a not negligible proportion (41.5%) of methods to setup the ground truth does not correspond to what standards considered as accepted reference values (in 19.5% the information was not even available). Despite annotation challenges (i.e., time consuming, big volume of data), it is essential to increase efforts to ensure systems measure their intended targets. Furthermore, considering expert variations in evaluation is crucial. Using strong agreement measures (e.g., Fleiss’ Kappa) on a subset ensures annotations aren’t dependent on a single annotator, facilitating the clinical implementation of CDSS.

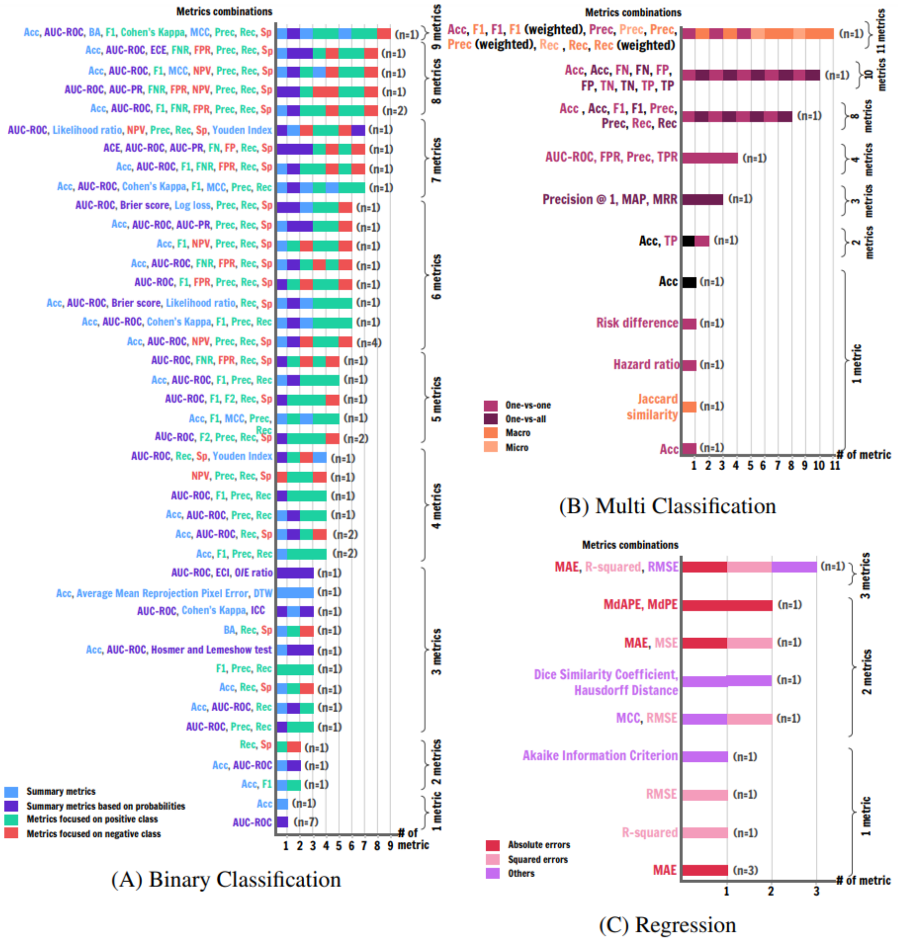

Figure 1. Metrics and combinations according to machine learning context, where for each combination the number of studies n is associated. A. In binary classification, four categories can be distinguished: summary metrics, summary metrics based on probabilities, metrics focusing on positive class, and metrics focusing on negative class. The metrics covered are: Area under ROC curve (AUC-ROC), Recall (Rec), Accuracy (Acc), Specificity (Sp), Precision (i.e, the metric) (Prec), F1-Score (F1), Negative Predictive Value (NPV), False Positive Rate (FPR), False Negative Rate (FNR), Cohen's Kappa, Matthew's Correlation Coefficients (MCC), Area under Precision-Recall curve (AUC-PR), F2-Score (F2), Youden Index, Balanced Accuracy (BA), Likelihood Ratio, Brier Score, Estimated Calibration Index (ECI), Expected Calibration Error (ECE), Absolute Calibration Error (ACE), Observed-to-expected Outcome Ratio (O/E ratio), False Positive (FP), False Negative (FN), Hosmer and Lemeshow test, Intraclass Correlation Coefficient (ICC), Average Mean Reprojection Pixel Error, Dynamic Time Warping (DTW). B. In multi-classification, "one-vs-one" (binary problem for each pair of class) and "one-vs-all" (binary problem for each class) was specified along with "macro" or "micro" computation. Metrics additionally encountered are: Jaccard Similarity, Risk Difference, Hazard Ratio, Precision @ 1, Mean Average Precision (MAP), Mean Reciprocal Rank (MRR). C. In regression, three categories are defined: absolute errors, squared errors, and others. Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R-squared, Hausdorff Distance, Dice Similarity Coefficient, Akaike Information Criterion, Median Percentage Error (MdPE), Median Absolute Percentage Error (MdAPE), Mean Square Error (MSE).

The observed metrics were mostly standards in machine learning. For classification, metrics include AUC-ROC, Accuracy, Recall, Specificity, and F1-Score. For regression, we often see MAE, RMSE, and R-Squared. It's notable that in binary classification, there's a tendency for larger combinations of metrics derived from confusion matrices. AUC-ROC is often used but may be insufficient since it does not directly provide the performance of the system after applying the threshold on outputted probabilities. Moreover, the under representation of metrics focusing on evaluating the performance for the negative class is a problem in properly assessing the trueness. Indeed, it underlines the underquantification of the risks associated with false positives (i.e., overdiagnosis) which is also crucial to obtain certification. We observed a weighty use of accuracy (i.e., the metric) while the nature of clinical problems are often unbalanced with an over-represented negative class. We didn't anticipate this, and it would be interesting to dig deeper to verify if balancing was done correctly before measuring. Another challenge is that the list of criteria we draw is an initial proposal that will evolve as consensual definition will be proposed. We hope that this work will stimulate discussion on the subject.

5. Conclusions

Focusing on evaluating Trueness, our findings reveal a gap between the definition of ground truth as expected by ISO standard and how it is defined in various studies. Moreover, we found that metrics used to assess trueness in binary classification mostly focus on performance on the positive class and tends to neglect performance in regard to the negative class. By bridging these gaps, authors may better pave the way for the successful integration of CDSS into clinical practice.

References

[1] Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, Rigby M, Scott PJ, Vehko T, Wong ZS, Georgiou A. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearbook of medical informatics. 2019 Aug;28(01):128-34.

[2] Li B, Qi P, Liu B, Di S, Liu J, Pei J, Yi J, Zhou B. Trustworthy AI: From principles to practices. ACM Computing Surveys. 2023 Jan 13;55(9):1-46. doi: 10.1145/3555803

[3] European Commission. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative Acts (AI Act). COM(2021) 206 final (European Commission, 2021).

[4] European Union. Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. 2017 O.J. (L 117) 1.

[5] Lekadir K, Osuala R, Gallin C, Lazrak N, Kushibar K, Tsakou G, Aussó S, Alberich LC, Marias K, Tsiknakis M, Colantonio S. FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. arXiv preprint arXiv:2109.09658. 2021 Sep 20.

[6] World Health Organization. Ethics and Governance of Artificial Intelligence for Health. Issued on 28 June 2021. Available online: https://iris.who.int/bitstream/handle/10665/341996/9789240029200-eng.pdf?sequence=1.

[7] Poiron A. Annexes of "How trueness of clinical decision support systems based on machine learning is assessed?" https://gitlab.com/alex.poiron/annexes-mie-2024.

[8] International Organization for Standardization: Geneva, Switzerland,. Accuracy (Trueness and Precision) of Measurement Methods and Results—Part 1: General Principles and Definitions. (ISO 5725-1: 2023).