How accurate do you want it? Defining minimum required accuracy for medical artificial intelligence

HOW ACCURATE DO YOU WANT IT? DEFINING MINIMUM REQUIRED ACCURACY FOR MEDICAL ARTIFICIAL INTELLIGENCE

Federico Sternini
USE-ME-D srl, I3P Politecnico di Torino
Corso Castelfidardo 30/a, Torino, Italy

Alice Ravizza
USE-ME-D srl, I3P Politecnico di Torino
Corso Castelfidardo 30/a, Torino, Italy

Federico Cabitza
Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca
Viale Sarca 336, 20126, Milano, *Italy*

ABSTRACT

Artificial intelligence (AI) is becoming a more and more common component of biomedical engineering solutions, and these latter systems are getting promising results in terms of diagnostic and prognostic accuracy. Medical AI (MAI) is then reaching the maturity level for its appropriate use in clinical practice, but to this end, its efficacy needs to be demonstrated first. Currently, this efficacy is proven in terms of the reported accuracy of the algorithm, especially for diagnostic tasks. But also in this case, how much accurate is "enough" accurate? To address this question means to define the minimum required accuracy for a system to be valid, that is fit to its intended use. To this aim, we propose a risk-based approach to the definition of adequate accuracy, in accordance with a risk-based regulatory classification. We investigated whether the current state of the art is already compliant with this standard-based approach, by performing a literature review in four application domains, one for each of the four risk classes we identified: the diagnosis of psoriasis, of knee osteoarthritis, the screening of breast cancer screening, and the detection of influenza outbreaks. The evaluation of the literature review highlighted that this approach is still not widely adopted, but that there is a partial presence of an implicit, conventional scheme that is similar to our proposal, especially in the high-impact literature. We also provide some guideline to assess the minimum required accuracy but also sheds light on the need for further official guidelines that ensure the wider application of the regulatory risk-based approach by the scholarly community of MAI.

KEYWORDS

accuracy, performance, validation, machine learning, medical artificial intelligence, standards

1. INTRODUCTION

Artificial Intelligence (AI) is acquiring increasing importance in all fields of technology and industry, including the healthcare domain. When AI is used to diagnose, prevent, monitor, predict, treat or alleviate diseases or injuries, it should be classified as a software as a medical device (FDA, 2019; GHTF, 2012) and we denote it as Medical AI (MAI) (Cabitza and Zeitoun, 2019). For the delicate and sensitive context in which is applied, MAI needs a regulatory path intended to prove compliance with the safety, effectiveness and quality requirements (i.e., verification), as defined by the relevant regulatory system and in many cases also prove some form of clinical effectiveness and actual benefit (i.e., validation). Regulations enacted worldwide all follow the World Health Organization recommendation (World Health Organization, 2011) and share a classification approach, which typically divides medical devices into different classes depending on the risk to the patient (European Parliament and European Council, 2017a, 2017b). Each class is then associated with different requirements, which increase as patient risk increases. The main goal of this approach is the efficient

distribution of resources from the competent authorities, where the most critical (and impactful) cases receive greater attention and resources.

Grounding on international standards and regulations, we decided to propose a classification of AI in terms of different levels of minimum required safety depending on the risk profile of the algorithm. We propose that those requirements should be defined for the whole software life cycle, from the training phase to the end of service, including all updates and further training phases. Our proposal is based on two regulatory references: the identification of the requirement for the pre-market assessment of the In Vitro Diagnostic devices drafted by the Global Harmonization Task Force (GHTC), and the framework for risk Categorization of Software as a Medical Device (SaMD) proposed by the International Medical Device Regulators Forum (IMDRF) (GHTF, 2008; IMDRF, 2014) on one hand, and the European Regulations 2017/745 and 2017/746, respectively on Medical Devices and InVitro Diagnostic medical devices, on the other hand (European Parliament and European Council, 2017a, 2017b). The choice of these references is backed up by consistency of them with the ones used by the Food and Drug Administration to propose a Regulatory Framework for modifications to MAI (FDA, 2019).

In this study, we propose an interpretation of the classification to analyze requirements for AI intended to complete diagnosis automatically (e.g., in screening ambit) or, more commonly, support the clinicians in their diagnostic tasks. The classification is defined as follows:

- I: Non-serious condition: A condition for which wrong or delayed diagnosis can cause minor injuries if any (e.g. acne diagnosis, psoriasis diagnosis).
- II: Serious condition: when wrong or delayed diagnosis can cause mild injuries to the patient (e.g. osteoarthritis).
- IIIa: Critical condition: when wrong or delayed diagnosis can cause death or severe injury (e.g. cancer, heart attack).
- IIIb: Public health threat: when wrong or delayed diagnosis can expose public health to harm (e.g. epidemic influenza, HIV).

In the specific context of AI embedded in diagnostic support systems, the safety is tightly correlated with the accuracy of the algorithm's output. All accuracy-related metrics are aimed at representing the extent a classifier makes mistakes, from a quantitative and probabilistic point of view. Specificity and sensitivity (and all their averages and combinations, like f- and g-scores) also express a qualitative, and yet crucial, aspect, by distinguishing between false negatives and false positive mistakes. However, to this respect, the most common approach is to weight both classes (i.e., positive and negative cases) equally, and this is clearly unrealistic in many cases. Moreover, to our knowledge, only one accuracy metrics takes the severity of condition into account, by weighting mistakes according to the complexity of the cases in the test set when this is annotated with this additional information (Cabitza and Campagner, 2019): however, this method has not yet found widespread application in medical ground truthing.

Our proposal takes into account the current regulation, which we here quote: "Devices shall achieve the performance intended by their manufacturer and shall be designed and manufactured in such a way that, during normal conditions of use, they are suitable for their intended purpose" (European Parliament and European Council, 2017a), and it grounds on an intuitive, if not obvious, notion: not all mistakes (of a classifier) are made equal, and the higher their potential impact on the patients' health, the lower the probability of their occurrence should be. In other words, we claim that: first, scholars should consider a classifier skill *acceptable* (and hence worthy of publication and dissemination) when it is better than previous already-released systems (being considered acceptable at their time); moreover, they should demand higher accuracy for more impactful errors and more serious conditions than for lighter ones. At the moment, we are not aware of similar proposals for the definition of the accuracy requirements for MAI.

In this paper, we will pursue then two research questions: first, we will report for a single prototypical health condition extracted from each one of the risk classes considered above, what the average accuracy of existing models is; in doing so, we can consider such an estimate a sort of lower bound threshold of acceptable accuracy for that condition (and, more loosely, for the corresponding risk class). Secondly, we will report whether the above idea to require higher accuracy for higher criticality is already an established and common practice in the specialist community of ML researchers.

To either the above aims, in light of the definition of the risk classes proposed above, we performed a concise, systematic review for one representative application of each risk class. The applications are the following ones:

- Non-serious condition: psoriasis;

- Serious condition: knee osteoarthrosis;
- Critical condition: Breast cancer;
- Public health threat: Influenza;

2. METHODS

To perform the literature review, four queries were executed on the article repository indexed by Pubmed on the 22nd of January 2020. The queries were designed to include only results published in English in the last five years. The following keywords were also used:
- "(((((artificial intelligence) OR deep learning) OR machine learning) AND diagnosis) AND [condition])", with [condition] being one of the prototypical conditions mentioned above.

2.1 Eligibility criteria and evaluation

All abstracts were manually reviewed, and the sources that were found not related to the specific disease or an AI application were excluded. Subsequently, the whole manuscript of the remaining articles was manually reviewed, and the articles found relevant to the research questions were included. In addition, the manuscripts were assessed in accordance with Meddev 2.7/1 rev 4 in terms of relevance to the research question (EUROPEAN COMMISSION, 2016). The assessment procedure requires to assign a score to each source, and it is composed of three parts. The first one regards the level of evidence reported in the manuscript, defined in accordance with the Oxford evidence scale (OCEBM Levels of Evidence Working Group, 2011); as shown in Table 1, the second part regards the impact factor of the journal, while the third contribution is determined by the application of the appraisal plan as described in Meddev 2.7/1 rev 4, as described in Table 2. The three sections maximum score are respectively 6, 6 and 18. The total result maximum score is 30. It should be noted that regarding the second element of the procedure described in Table 2, the required performance should not be defined by simply comparing similar studies of similar applications.

Table 1. Scores based on the evidence level

Description	Score
Systematic review of cross sectional studies with consistently applied reference standard and blinding	6
Individual cross sectional studies with consistently applied reference standard and blinding	4
Non-consecutive studies, or studies without consistently applied reference standards	2
Case-control studies, or "poor or non-independent reference standard	1
Mechanism-based reasoning	0

Table 2. Score on the base of the Meddev appraisal plan

Description	Examples	Score
To what extent are the data generated representative of the device under evaluatior	The device is an AI whose output is a diagnosis or a diagnostic suggestion	6
	The device is an AI whose output is additional information to aid the diagnosis	3
	Other devices	Not Applicable
What aspects are covered?	Required performances	8
	Performances	4
	Other aspects	Not Applicable
Are the data relevant to the intended purpose of the device or to claims about the device?	Searched disease	4
	concerns specific models/ sizes/ settings, or concerns specific aspects of the intended purpose or of claims	0
	Other diseases	Not Applicable
If the data are relevant to specific aspects of the intended purpose or claims, are they relevant to a specific	Specific type OR severity of the medical condition	2

- type and severity of the medical condition?

Specific type AND severity of the medical condition	0

Finally, the scores of the different categories are evaluated to understand if the found evidence is compatible with our proposal or not.

In addition, during all the literature review process, information regarding the performance of the applications in terms of accuracy, AUC (c-statistics), f-score, sensitivity and sensibility have been collected. These values have then been used for a twofold aim coherent with our two research questions: first to propose a level of minimum necessary accuracy for each risk level; then to understand if the literature supports our proposal. To this aim, we adopt the following approach: the average value of the metric per application is calculated, and then, if the relationship of the average values with the average value of the other applications is consistent with the proposed approach (i.e. the accuracy is higher as the risk level is higher), the papers reporting the metric are considered as supporting our framework; in the case of metric values not consistent with the proposed scheme, the negative evidence is calculated as the average of the scores of studies presenting those metrics. In addition to data related to Meddev evaluation and accuracy data, we analysed how many of the papers compared the MAI accuracy with accuracy performed by human operators

As a last task, we also collected the qualitative terms associated with each paper to evaluate how these attributes are related to different accuracy levels. To assign a score to each qualitative value, three coders rated all of the qualitative attributes associated with the judgments of performance found in the reviewed articles on a 4-value scale, with no neutral code to minimize central-tendency bias: strong negative, weak negative, weak positive and strong positive.

3. RESULTS

The research performed on PubMed identified 124 articles. After the abstract and full-text evaluation, 59 articles were identified as applicable, divided as follows between the different applications:

- Psoriasis diagnosis: 7 papers;
- Knee Osteoarthritis diagnosis: 25 papers;
- Breast cancer screening: 17 papers;
- Influenza detection: 10 papers;

All of the papers were evaluated, and were associated with at least 12 out of 30 points as their evidence score; no article got a score higher than 25 out of 30. The average evidence score was 18. A vast majority of studies were not compatible with our proposal (52 vs 19), while the evidence scores were almost evenly matched with a slight primacy of the studies backing up our proposal (18.5 vs 17.9). The details of all the papers evaluated are listed in the appendix to this paper, which is available at shorturl.at/GKO19[1].

52 out of 59 papers (88%) included at least one value of the accuracy metrics listed above, thus leading to the collection of 22 values for the accuracy, 28 AUC values, 12 f-score and 21 values of both sensitivity and specificity. The average scores are reported in table 3. Moreover, 13 studies out of 59 (22%) compared the MAI accuracy with the accuracy obtained by human operators.

Table 3. Average accuracy metrics grouped by application

Application	Accuracy (Mean ± St.Dev.)	AUC (Mean ± St.Dev.)	f-score (Mean ± St.Dev.)	Sensitivity (Mean ± St.Dev.)	Specificity (Mean ± St.Dev.)
Psoriasis diagnosis	0.86 ± 0.21	0.87 ± 0.00	0.59 ± 0.00	0.75 ± 0.10	0.70 ± 0.00
Osteoarthritis diagnosis	0.77 ± 0.10	0.83 ± 0.09	0.78 ± 0.04	0.81 ± 0.11	0.80 ± 0.11
Breast cancer screening	0.76 ± 0.28	0.84 ± 0.11	0.88 ± 0.03	0.87 ± 0.08	0.74 ± 0.31
Influenza detection	0.87 ± 0.00	0.89 ± 0.05	0.84 ± 0.08	0.93 ± 0.04	0.90 ± 0.09

[1] The original URL is: https://drive.google.com/file/d/1ZXf2Mr-nbvhY-JPTvquV0vje013MHkBI/view

Breast cancer screening and Influenza detection	0.78	±	0.26	0.85	±	0.11	0.86	±	0.07	0.88	±	0.08	0.78	±	0.29

Regarding the qualitative attributes characterizing the MAI performance, three coders rated 26 expressions and their agreement, computed in terms of Krippendorff's Alpha, was found "substantial" (0.62) (Landis and Koch, 1977); this allowed us to proceed with majority voting and assign each accuracy score with an appreciation level. This classification yielded to 19% strong positive appreciations, 68% weak positive ones, 17% weak negative ones, and no strong negative ones. We anticipate here that this latter result was expected, as models that perform so badly to be connotated in strongly negative terms hardly are published in impacted journals (by either file-drawer effect or review filtering). Thus, the performance of the vast majority (more than two thirds) of the predictive models reviewed was characterized in weakly positive terms by their developers or authors: this class was associated with average and minimum scores listed in Table 4.

Table 4. Average, median and minimum accuracy of models characterized by weakly positive terms

	Psoriasis diagnosis	Knee osteoarthrosis diagnosis	Breast cancer screening	Influenza detection	Breast cancer screening and influenza detection	Global
Average	48.75%	79.17%	94.55%	86.70%	92.59%	82.07%
Median	48.75%	82.50%	97.91%	86.70%	92.71%	86.35%
Minimum	48.75%	69.00%	87.50%	86.70%	86.70%	48.75%

Table 5. Evaluation of the qualitative judgements expressed by the authors. The values from 1 to 4 correspond to the grades Strong Negative, Weak Negative, Weak Positive, Strong positive

Judgement of performance	Coder 1	Coder 2	Coder 3	Majority vote
Highest accuracy	4	4	4	4
Appreciable	3	3	3	3
Good agreement	3	3	3	3
Outperforms human	4	4	4	4
Considerably high	4	4	4	4
Robust	3	3	3	3
Success	3	3	3	3
relatively high	3	3	2	3
Satisfactory	3	3	2	3
Fair	3	3	2	3
High	3	4	3	3
Promising	2	2	1	2
Good	3	3	3	3
Reliable	3	3	3	3
Encouraging	2	2	2	2
well-performing	3	3	2	3
Accurate	3	3	3	3
high level accuracy	4	4	4	4
potential effectiveness	3	3	2	3
comparable with experts	4	4	4	4
Acceptable	3	3	3	3
Accurately	3	3	3	3

4. DISCUSSION

As a first consideration, let us consider a parameter that was not reported in the Results Section but nevertheless can inform the interpretation of the findings: the percentage of papers that included an explicit requirement of the accuracy beyond the simple comparison with the performance of similar applications. This parameter was not used to evaluate our proposal because the requirement definition is part of the Meddev: therefore, it would have inflated our results on the positive side. The percentage of papers that explicitly declared the requirement in the context of the psoriasis diagnosis was 14.3%, while in the context of osteoarthritis was 28% and in the context of breast cancer screening 29.4%. None of the papers concerned with flu detection described the required accuracy for the system, probably due to the difficulty and novelty of the application domain. Indeed, while the other applications are consolidated in the medical practice and are usually performed by specialists, the detection of a flu outbreak is not an activity that is part of the common clinical practice completed by single physicians. In addition, but not less importantly, many of the current flu applications are designed to early detect flu outbreaks by means that are out of the usual scope of routine clinical evaluations, like the use of social media and search query analysis. These peculiarities make understanding what the adequate threshold could be of performance acceptability difficult for these applications.

Another parameter that shall be considered is the percentage of studies that involved the comparison with the human operator. Even in this case, the percentage increases as also the risk level increases (psoriasis diagnosis: 14.3%; osteoarthrosis diagnosis: 24%; breast cancer screening: 35.3%), but this also drops to zero for flu detection. All the possible reasons identified above can also be used in this case: the proposed systems are not intended to give a second opinion to the physician but are intended to detect situations that are defined as dangerous for the public health to epidemiologists and health policymakers.

An additional parameter that could be considered is the percentage of studies that compared the results with already published literature. In this case, the percentage decreases as the risk level increases (psoriasis diagnosis: 57.14%, osteoarthrosis diagnosis: 52%, breast cancer screening: 35,29%, flu detection: 30%). Even if this trend could seem to be related to the risk level of the different applications, we decided to exclude this term from further analysis due to the high number of factors that could affect the availability of a literature evaluation (e.g. novelty of the technology) and that are out of the scope of this review.

For the above reasons, in what follows we will focus on the first three risk levels (I, II, IIIa), as those concerning individual clinical cases (indeed, we recall here that the public health case is associated with a higher magnitude of risk, IIIb, not for the disease criticality per se but for the multitude of patients that can potentially be affected by it). In this case, accuracy and AUC present an opposite trend with respect to our proposal: the average performance reported in the reviewed studies decrease as the risk class these studies belong to increase, with the last two risk levels being almost evenly matched. In fact, the only two metrics that are consistent with our proposal are the sensitivity (which is so in regard to all four risk levels) and the f-score. However, as counterintuitively as it might seem, these findings can be read as a confirmation of the reasonability of our framework and as an indication of what difficulties the scientific community is currently facing. In fact, we recall that sensitivity regards the capability of the MAI to avoid false-negative errors: intuitively, the more serious the condition, the higher impact the consequences of not detecting and treating the condition timely, and the harm for the patient. This intuition must have passed from the medical community to the MAI practitioners, who, in trying to optimize their predictive models, must have favoured sensitivity for potentially serious conditions. This is also mirrored by the finding of the f-score, which is the (harmonic) mean of the model's sensitivity and its capability not to raise false positive alarms (cf. the positive predictive value of the model).

The opposite trend for the accuracy metrics does not come totally unexpected: lower risk conditions are such also for their higher recognizability and their relationship with a smaller number of physiological states and causal links; this could be mirrored by perceptual tasks that are easier to automate and smaller predictor spaces to be computationally represented. State of the art is more than adequate to detect their signs and identify powerful predictive features. The same does not hold for most serious and complex health conditions.

Also, the quantitative analysis of the reviewed studies confirms this reading of the results: Whilst more studies were not compatible with our proposal (52 vs 19), the evidence score associated with the compatible studies is slightly higher (18.5vs 17.9). Intuitively this means that the higher-impact resources are already more aligned with our proposal, while more resources contradict it but on (relatively) lower-value sources. Therefore, for the moment, we can claim that the majority of recent studies on diagnostic MAI do not implicitly adopt our

approach to self-validate their results, but the most impacted (and higher-value) contributions show a tendency to its adoption: that notwithstanding, validation is commonly based on the comparison with the performance of similar previous algorithms or the subjective evaluation of the authors, without considering the risk class of the intended use.

From the analysis of the qualitative terms, we here focus on the accuracy values that we associated with the weak positive scale. The intuitive idea is that the average value of the accuracy reported by the studies falling in this category would provide a threshold for acceptability that is based on the implicit consensus of the scholarly community. Moreover, the weak positive class was the more equally distributed among all risk classes (Class I: 12,5%, Class II: 50%, Class IIIa: 25%, Class IIIb: 12,5%), thus providing mean scores that are internally more representative. The total average across all risk classes, as reported in Table 4, is 82%, which is nevertheless a pessimistic (low) estimate, also for the minimum extreme that is even lower than the chance guess (i.e., 50%). The average found for each class also provides an intuitive criterion for the specific health conditions we chose to focus on, as it represents the average performance of comparable applications whose performance was found to be at least weakly positive by their developers. Similar consideration should be done for each single health condition, as reasoning for risk class would likely provide too coarse-grained indications, for the high heterogeneity of diagnostic class in the same risk class. All things considered, a gross, but easy-to-remember, rule of thumb to set a minimum acceptable for diagnostic MAI is what we call the "die-roll" threshold.: Every side of dice has one sixth probability of getting out (ca. 17%); this could be the acceptable error rate for decent decision support, which also finds an interesting parallel with recent estimates of human diagnostic error, that is, 13%-15%, (Graber, 2013). The die-roll threshold that is substantially confirmed by our bibliographic study would then be more or less a "psychological" level, set at approximately 83% (accuracy).

5. CONCLUSION

In this work, we have proposed a general approach for the definition of requirements of minimum accuracy for AI applications in medicine: a risk-based approach with a classification of the application risk that is consistent with the current regulatory framework worldwide. We performed a literature survey to understand if this approach is already applied in the scientific community of those who develop and deploy medical AI systems.

The above literature survey was completed in four application domains of increasing severity and impact to be loosely representative of the whole medical domain. The literature evaluation showed that, for the moment, the most used approach is an "opportunistic" approach that does not consider the risk of the application but relies on the judgement of the authors or comparison with similar algorithms. Even though, the literature analysis suggests that a similar validation approach is getting wider and wider application, especially in the most impacted specialist literature, and in regard to the risk of committing a false negative mistake, that is the sensitivity metric.

After the collection of all the accuracy score for the different applications, we propose these values as state-of-the-art baseline and acceptability threshold for future studies that are in the scope of our literature review.

This study also suggests that providing a clear guideline may encourage future studies to take regulatory requirements more seriously, thus setting higher expectations for those models that are intended to treat or diagnose more critical conditions.

REFERENCES

Cabitza, F., Campagner, A., 2019. Who wants accurate models? Arguing for a different metrics to take classification models seriously. ArXiv191009246 Cs Stat.

Cabitza, F., Zeitoun, J.-D., 2019. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. 2019 7, 1.

EUROPEAN COMMISSION, 2016. CLINICAL EVALUATION: A GUIDE FOR MANUFACTURERS AND NOTIFIED BODIES UNDER DIRECTIVES 93/42/EEC and 90/385/EEC, MEDDEV 2.7/1 revision 4.

European Parliament, European Council, 2017a. 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 (Text with EEA relevance.), OJ L.

European Parliament, European Council, 2017b. Regulation (EU) 2017/746 of the European Parliament and of the Council of 5 April 2017 on in vitro diagnostic medical devices and repealing Directive 98/79/EC and Commission Decision 2010/227/EU (Text with EEA relevance.), OJ L.

FDA, 2019. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback.

GHTF, 2008. Principles of In Vitro Diagnostics (IVD) Medical Devices Classification.

GHTF, 2012. Definition of the Terms 'Medical Device' and 'In Vitro Diagnostic (IVD) Medical Device.'

Graber, M.L., 2013. The incidence of diagnostic error in medicine. BMJ Qual. Saf. 22, ii21–ii27.

IMDRF, 2014. "Software as a Medical Device": Possible Framework for Risk Categorization and Corresponding Considerations.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

OCEBM Levels of Evidence Working Group, 2011. The Oxford 2011 Levels of Evidence.

World Health Organization (Ed.), 2011. Development of medical device policies, WHO medical device technical series. World Health Organization, Geneva, Switzerland.