The need to separate the wheat from the chaff in medical informatics

Federico Cabitza\textsuperscript{a}, Andrea Campagner\textsuperscript{a}

\textsuperscript{a}DISCo, University of Milano-Bicocca, viale Sarca 336, Milano, 20126, Italy

Abstract

This editorial aims to contribute to the current debate about the quality of studies that apply machine learning (ML) methodologies to medical data to extract value from them and provide clinicians with viable and useful tools supporting everyday care practices. We propose a practical checklist to help authors to self-assess the quality of their contribution and to help reviewers to recognize and appreciate high-quality medical ML studies by distinguishing them from the mere application of ML techniques to medical data.

Keywords: Medical Artificial Intelligence, Machine Learning, Checklist, Quality auditing

As widely known, machine learning (ML) models are beginning to demonstrate early successes in clinical applications \cite{1, 2, 3}. Studies that compare the performance of these models and human physicians found that models allegedly perform equally well in many diagnostic and prognostic tasks \cite{4, 5}. However, relatively few studies present externally validated results \cite{6, 7, 8}, and most of them failed to adhere to minimal reporting standards \cite{9, 10}. In this respect, poor reporting is one of the main factors preventing studies from being replicated in other settings \cite{11}, which undermines the interpretation of the scores that authors report to estimate the diagnostic accuracy of the model on unseen data.

The “reproducibility crisis”, which some observers report affecting biomedical science \cite{12} at an increasing extent, also affects also medical informatics \cite{13}, artificial intelligence \cite{14} and its application to medicine \cite{15}. To quote a oft-cited work by Ioannidis \cite{16}, which could be seen as a precursor to the current debate on reproducibility in science and medicine, we know that “most published accuracy scores are false” or, more prosaically, “most published studies applying ML techniques to medicine are simply not valid”.

Preprint submitted to IJMEDI August 31, 2021
This assertion looks like the notorious elephant in the room \([17, 18]\) of medical informatics that few people want to escort out of the room.

The sheer truth is that practicing “Medical ML” is different from merely applying ML to medical data. Applying ML to medical data is relatively easy, once medical data are available. And they are: an increasing number of medical datasets have been made available to researchers and shared in public repositories in recent times: for example, HealthData\(^1\) is a U.S. site that collects data from agencies from the U.S. Department of Health and Human Services as well as other centers and counts to date more than 4,500 datasets that can be used to train ML models on disparate medical tasks; MIMIC-III \([19]\) is a freely accessible database with more than 60,000 intensive care unit admissions, that has been mentioned in more than 1,400 articles indexed in Scopus; OpenfMRI \([20]\) includes 95 datasets of magnetic resonance imaging (MRI) from more than 3,000 subjects, while Deep Lesion \([21]\) is a U.S. National Institutes of Health initiative to make more than 32,000 lesions in CT images, from 4,000 unique patients, available to foster research, better diagnostics and training. Moreover, on Kaggle and Healthcare.ai, which are popular sites visited by thousands of data science practitioners every day, ML researchers can find countless datasets that make training a ML model to predict some target variable a child’s play. However, few of these datasets would be considered high quality from a clinical point of view \([6, 22]\) and very seldom can we know how these data were produced (e.g., by involving how many experts, what their certification is, the conditions in which they performed their ratings), as a guarantee of their reliability at face level \([18]\).

Thus, mere data availability cannot be a sufficient condition to perform valid research in the field of medical ML: being at the intersection between data science, computer science and medicine, this subfield differs from the mere application of ML techniques to medical data. Medical ML is programmatically aimed at developing tools that medical doctors, nurses and other healthcare practitioners can use in their daily practice to improve the appropriateness, safety and effectiveness of their decisions, and ultimately the health outcomes of their patients \([23]\): thus, actual use and assessment are part and parcel of medical ML. This ambitious objective justifies efforts for which data scientists, who are increasingly focused on developing methods and techniques that apply to “big data” (which are impossible to vet for

\(^1\)http://www.healthdata.org/
actual reliability in order to gain marginal, if statistically significant at all, improvements over the state of the art [24]), are not usually interested in devoting themselves to.

Conversely, practicing medical ML often means dealing with relatively small datasets [25] (much smaller than what would be required to produce generalizable models using deep learning, or other equally complex approaches [26]), which are collected from real-world practice by vetting them for clinical meaning, and pose challenges [27] that are hardly, if ever, addressed in computer science laboratories: observer variability [28]; pre-analytical, analytical [29] and biological variability [30]; class imbalance [31]; small cardinality [32] (hence the consequent risk of overfitting); relatively high missing rate [33]; feature collinearity [34]; and any heterogeneity that may break the assumption of independence and identical distribution of data [35] or affect the variability of results [36].

Under the pressure of funding policies and assessment exercises that foster the “publish or perish” environment, medical informatics journals, and the IJMEDI is no exception, are flooded with contributions that do not address any of the problems that were previously mentioned, and that mechanically apply procedures which, by their nature, lend themselves to the growing trend toward automation (cf. autoML [37]). The same situation occurs in more technology- and application-oriented journals, which face similar difficulties in curbing a vast amount of articles that communities of peers find increasingly difficult to filter out, contributing to unintentionally creating precedents in the literature, which inspire works of similar superficiality [38].

As public opinion and many practitioners seem to be dazzled by discourses regarding the quality of instruments that do not extend beyond reports on their theoretical error rate (often not considering class imbalance or separating training data from validation data) [38], some scientific societies have recently suggested more sensible guidelines for assessing the quality, validity and usefulness of certain instruments in the medical field, and report on them. Recent collaborative efforts for the definition of guidelines on the development and reporting of Medical AI systems, see also [39], include the SPIRIT-AI [40] and CONSORT-AI [41] for the design and reporting of clinical trials involving AI and ML systems, the MI-CLAIM [42] checklist for Medical AI, the WHO/ITU ML4H auditing framework [43, 44] for artificial intelligence in healthcare, the PROBAST tool [45] to assess the bias and applicability of prediction models, or the TRIPOD statement [46] for reporting their main characteristics. To some extent, the availability of mul-
ultiple guidelines, as well as their long production time (as of the writing of this manuscript the TRIPOD-AI extension, which was announced in 2019 [41, 47], as well as the STARD-AI reporting guidelines [48], have not yet been officially published), indicate the difficulty of convening on a minimum set of data that must be reported to make ML studies reproducible and their results reliable.

In the light of the above partly overlapping and competing standards, we at the IJMEDI have considered the progress made by the recent proposals by the Journal of the Medical Informatics Association (JAMIA) [49], and by the BMJ Health & Care Informatics [25], a huge step forward, especially for their practical value. We consider these contributions powerful tools to improve the quality of ML studies, as a positive side effect of improving the reporting practices of their authors, and a way to disseminate good development practices. For this reason, we took inspiration from these relevant contributions to propose an even more assessment-oriented checklist: the IJMEDI checklist for assessment of medical artificial intelligence based on machine learning; in this tool some aspects are made even more explicit and detailed than in similar proposals, the aspects that we deem more relevant to allow our associate editors and reviewers to discriminate between high-quality contributions and manuscripts that should be rejected because of failing to meet the high standards of a journal that is so committed to the sound evaluation of computational systems in healthcare settings.

The following 30-item checklist, organized in 6 phases according to the CRISP-DM methodology [50], can be considered a practical guideline, for both reviewers and authors, to qualitatively assess the methodological soundness of a medical ML contribution and the reproducibility of its results. In the following list, each item represents a requirement and is associated with three possible options, for both authors (Not Applicable, Not Addressed – No, Addressed – Yes); and reviewers (Adequately addressed – OK, sufficient but improvable, minor revision needed – mR), inadequately addressed, major revision needed – MR). Items for which mR has been proposed can be interpreted as opportunities for due improvement; by contrast, items for which a MR has been proposed should be mandatorily addressed or considered as good reasons for rejecting the manuscript, and particularly so in the case the involved item is considered high priority (in bold) or if many of the requirements were considered inadequately addressed. Authors can help editors and reviewers by attaching the checklist to their manuscript and indicating which items have been addressed and which items are missing (and why).
The IJMEDI checklist for assessment of medical AI

<table>
<thead>
<tr>
<th>Requirement</th>
<th>Authors</th>
<th>Reviewers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td><strong>Problem Understanding</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Is the study population described, also in terms of inclusion/exclusion criteria (e.g., patients older than 18 tested for COVID-19; all inpatients hospitalized for 24 or more hours)? §</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>2. Is the study design described? (e.g., retrospective, prospective, cross-sectional [51], observational, randomized control trial [52]) §</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>3. Is the study setting described? (e.g., teaching tertiary hospital; primary care ambulatory, nursing home, medical laboratory, R&amp;D laboratory) §</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>4. Is the source of data described? (e.g., electronic specialty registry; laboratory information system; electronic health record; picture archiving and communication system) §</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>5. Is the medical task reported? (e.g., diagnostic detection, diagnostic characterization, diagnostic staging, prognosis on which endpoint), event prediction, risk stratification, anatomical structure segmentation, treatment selection and planning, monitoring) §</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>6. Is the data collection process described, also in terms of setting-specific data collection strategies (e.g. whether body temperatures are measured only in the morning; whether some blood tests are performed only in light of a specific diagnostic hypothesis)? Any consideration about data quality is appreciated, e.g., in regard to completeness, plausibility, and robustness with respect to upcoding or downcoding practices</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td><strong>Data Understanding</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Are the subject demographics described in terms of</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1. average age (mean or median);</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. age variability (standard deviation (SD) or inter-quartile range (IQR)):</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. gender breakdown (e.g., 55% female, 44% male, 1% not reported); §</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. main comorbidities;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. ethnic group (e.g., Native American, Asian, South East Asian, African, African American, Hispanic, Native Hawaiian or Other Pacific Islander, European or American White);</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. socioeconomic status?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. If the task is supervised, is the gold standard described? (e.g., &quot;100 manually annotated clinical notes and pain scores recorded in EHR, Death, re-admission and International Classification of Disease (ICD) codes in discharge letters&quot;). In particular, the authors should describe the process of ground truthing described in terms of:</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>1. Number of annotators (raters) producing the labels;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Their profession and expertise (e.g., years from specialization or graduation);</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Particular instructions given to annotators for quality control (e.g., which data were discarded and why);</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Inter-rater agreement score (e.g., Alpha [53], Kappa [54], Rho [17]);</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Labelling technique (e.g., majority voting, Delphi method [55], consensus iteration).</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
9. In the case of tabular data, are the features described (also in regard to how they were used in the model in terms of categories or transformation)? This description should be done for all, or, in the case that the features exceed 20, for a significant subset of the most predictive features in the following terms: name, short description, type (nominal, ordinal, continuous), and

1. If continuous: unit of measure, range (min, max), mean and standard deviation (or median and IQR). Violin plots of some relevant continuous features are appreciated. If data are hematochemical parameters, also mention the brand and model of the analyzer equipment.
2. If nominal, all codes/values and their distribution. Feature transformation (e.g. one-hot encoding) should be reported if applied. Any terminology standard should be explicitly mentioned (e.g., LOINC [56], ICD-11 [57], SNOMED [58]) if applied.

10. Is outlier detection and analysis performed and reported? If the answer is yes, the definition of an outlier should be given and the techniques applied to manage outliers should be described (e.g., removal through application of an Isolation Forest model).

11. Is missing-value management described? This description should be reported in the following terms:

   1. The missing rate for each feature should be reported;
   2. The technique of imputation, if any, should be described, and reasons for its choice should be given. If the missing rate is higher than 10%, a reflection about the impact on the performance of a technique with respect to others would be appreciable [59].

12. Is feature pre-processing performed and described? This description should be reported in terms of scaling transformations (e.g., normalization, standardization, log-transformation) or discretization procedures applied to continuous features, and encoding of categorical or ordinal variables (e.g., one-hot encoding, ordinal encoding).

13. Is data imbalance analysis and adjustment performed and reported? The authors should describe any imbalance in the data distribution, both in regard to the target (e.g. only 10% of the patients were affected by a given disease); and in regard to important predictive features (e.g. female patients accounted for less than 10% of the total cases). The authors should also report about any technique (if any) applied to adjust the above mentioned imbalances (e.g. under- or over-sampling, SMOTE).

14. Is the model task reported? (e.g., binary classification, multi-class classification, multi-label classification, ordinal regression, continuous regression, clustering, dimensionality reduction, segmentation) §

15. Is the model output specified? (e.g., disease positivity probability score, probability of infection within 5 days, postoperative 3-month pain scores) §

16. Is the model architecture or type described? (e.g., SVM, Random Forest, Boosting, Logistic Regression, Nearest Neighbors, Convolutional Neural Network)

17. Is the data splitting [60] described (e.g., no data splitting, k-fold cross-validation (CV); nested k-fold CV; repeated CV; bootstrap validation; leave-one-out CV; 80%/10%/10% train/validation/test)? In the case of data splitting, the authors must explicitly state that splitting was performed before any pre-processing steps (e.g. normalization, standardization, missing value imputation, feature selection) or model construction steps (training, hyper-parameter optimization), so to avoid data leakage [61] and overfitting.
18. Is the model training and selection described? In particular, the training procedure, hyper-parameter optimization or model selection should be described in terms of:

1. Range of hyper-parameters [62];
2. Method used to select the best hyper-parameter configuration (e.g., Hyper-parameter selection was performed through nested k-fold CV based grid search);
3. Full specification of the hyper-parameters used to generate results [64];
4. Procedure (if any) to limit over-fitting, in particular as related to the sample size [25].

19. (classification models) Is the model calibration described? If the answer is yes, the Brier score should be reported, and a calibration plot should be presented [63].

20. Is the internal/external model validation procedure described [64] (e.g., internal 10-fold CV, time-based cross-validation)? The authors should explicitly specify that the sets have been split before normalization, standardization and imputation, to avoid data leakage [61] (also refer to item 17 of this guideline). If possible, the authors should also comment on the adequacy of the available sample size for model training and validation [65, 25]. Moreover, the authors should try to choose the test set so that it is the most diverse with respect to the remainder of the sample [66] (w.r.t. some multivariate similarity function) and how this choice relates to conservative (and lower-bound) estimates of the model’s accuracy (and performance).

21. Has the model been externally validated [67]? If the answer is yes, the characteristics of the external validation set(s) should be described. For instance, the authors could comment about the heterogeneity of the data with respect to the training set (e.g., degree of correspondence $\Psi$ [66], Data Representativeness Criterion [65]) and the cardinality of the external sample [69]. If the performance on external datasets is found to be comparable with (or better than) that on training and internal datasets, the authors should provide some explanatory conjectures for why this happened (e.g., high heterogeneity of the training set, high homogeneity of the external dataset).
22. Are the main error-based metrics used?

1. a. Classification performance should be reported in terms of: Accuracy, Balanced accuracy, Specificity, Sensitivity (recall), Area Under the Curve (if the positive condition is extremely rare - as in case of stroke events - authors could consider the “Area under the Precision-Recall Curve” [70]). Optionally also in terms of: positive and negative predictive value, F1 score, Matthew coefficient [71], F score of sensitivity and specificity, the full confusion matrix, Hamming Loss (for multi-label classification), Jaccard Index (for multi-label classification).

2. Regression performance should be reported in terms of: \( R^2 \); Mean Absolute Error (MAE); Root Mean Square Error (RMSE); Mean Absolute Percentage Error (MAPE) or the Ratio between MAE (or RMSE) and SD (of the target).

3. Clustering performance should be reported in terms of: External validation metrics (e.g. mutual information, purity, Rand index), when ground truth labels are available, and Internal validation metrics (e.g. Davies-Bouldin index, Silhouette index, Homogeneity). The reported results of internal validation metrics should be discussed [72].

4. Image segmentation performance, depending on the specific task, should be reported in terms of metrics like [73]: accuracy-based metrics (e.g. Pixel accuracy, Jaccard Index, Dice Coefficient), distance-based metrics (e.g. mean absolute, or maximum difference), or area-based metrics (e.g. true positive fraction, true negative fraction, false positive fraction, false negative fraction).


The above estimates should be expressed, whenever possible, with their 95% (or 90%) confidence intervals (CI), or with other indicators of variability [56], with respect to the evaluation metrics reported. In this case, the authors should report which methods were applied for the computation of the confidence intervals (e.g. whether k-fold CV or bootstrap was applied, normal approximation). When comparing multiple models, the authors should discuss the statistical significance of the observed differences [75] (e.g. through CI comparisons, or hypothesis testing). When comparing multiple regression models, a Taylor diagram [76] could be reported and discussed.

23. Are some relevant errors described? The authors should describe the characteristic of some noteworthy classification errors or cases for which the regression prediction was much higher (> 2x) than the MAE. If these cases represent statistical outliers for some covariates, the authors should comment on that. To detect relevant cases, the authors could focus on those cases on which the inter-rater agreement (either re ground truth or by comparing human vs. model’s performance) is lowest.

24. Is the target user indicated? (e.g., clinician, radiologist, hospital management team, insurance company, patients) §

25. (classification models) Is the utility of the model discussed? The authors should report the performance of a baseline model (e.g., logistic regression, Naive Bayes). Additionally, the authors could report the Net Benefit [74] or similar metrics and present utility curves [75]. In particular, the authors are encouraged to discuss the selection of appropriate risk thresholds [79]; the relative value of benefits (true positives/negatives) and harms (false positives/negatives); and the clinical utility of the proposed models [80].
26. Is information regarding model interpretability and explainability available (e.g. feature importance, interpretable surrogate models, information about the model parameters)? Claims of “high” or “adequate” model interpretability (e.g., by means of visual aids like decision trees, Variable Importance Plots or Shapley Additive Explanations Plots (SHAP) [81]) or model causability [82] should always be supported by some user study, even qualitative or questionnaires-based [83]. In the case surrogate models were applied, the authors should report about their fidelity [84, 85].

27. Is there any discussion regarding model fairness, ethical concerns or risks of bias [25, 86] (for a list of clinically relevant biases, refer to [87])? If possible, the authors should report the model performance stratified for particularly relevant population strata [88] (e.g., model performance on male vs female subjects, or on minority groups).

28. Is any point made about the environmental sustainability of the model, or about the carbon footprint [89], of either the training phase or inference phase (use) of the model? If the answer is yes, then such a footprint should be expressed in terms of carbon dioxide equivalent (CO₂eq) and details about the estimation method should be given. Any efforts to this end will be appreciated, including those based on tools available online[2] as well as any attempts to popularise this concept, e.g. through equivalences with the consumption of everyday devices such as smartphones or kilometres travelled by a fossil-fuelled car[3].

29. Is code and data shared with the community [62, 90]? § If not, are reasons given? If code and data are shared, institutional repositories such as Zenodo should be preferred to private-owned repositories (arxiv, GitHub). If code is shared, specification of dependencies should be reported and a clear distinction between training code and evaluation code should be made. The authors should also state whether the developed system, either as a sand-box or as fully-operating system, has been made freely accessible on the Web.

30. Is the system already adopted in daily practice? If the answer is yes, the authors should report on where (setting name) and since when. Moreover, appreciated additions would regard: the description on the digitized workflow integrating the system, any comment about the level of use [29], a qualitative assessment of the level of efficacy of the system’s contribution to the clinical process (e.g. [91, 92]); any comment about the technical and staff training effort actually required [29]. If the answer is no, the authors should be explicit in regard to the point in the clinical workflow where the ML model should be applied, possibly using standard notation (e.g., BPMN). Moreover, the authors should also propose an assessment of the technology readiness of the described system with explicit reference to the Technology Readiness Level framework[4] or to any adaptation of this framework to the AI/ML domain [93]. In either above cases (yes/no), the authors should report about the procedures (if any) for performance monitoring, model maintenance and updating [94].

Table 1: Checklist for assessment of requirements and recommendations for sound medical ML contributions to the existing literature. NA: not applicable; mR: minor revisions needed; MR: major revisions needed. Items in bold indicate priority aspects to be considered. Items denoted with a § symbol are directly inspired by the MINIMAR guideline [49]. The section names for the checklist items are directly inspired by the CRISP-DM framework [50].

To download a copy of the above checklist, see: https://zenodo.org/record/4835800#.YLDlaaGxVPY

---

2https://mlco2.github.io/impact/
3https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator
4Technology readiness levels (TRL) - Extract from Part 19 - Commission Decision C (2014) 4995
Acknowledgments

Federico Cabitza is grateful to the IJMEDI Editor-in-chief, Heimar De Fatima Marin for her continuous support and encouragement in writing this editorial. The authors are also grateful to the members of the editorial board for their comments on the draft of this paper, and especially to Riccardo Bellazzi, for the valuable suggestions made during the review process.

Summary Points

What was already known

• Recent studies reported on common pitfalls and challenges in the development of medical ML systems, highlighting their general lack of reproducibility and reliability;

• Several proposals for reporting guidelines have been proposed in the literature to address these challenges and improve the quality of ML studies aimed at supporting clinical practice;

What does this study adds to our knowledge

• We propose a comprehensive checklist for the self-assessment and evaluation of medical ML papers, encompassing a set of 30 requirements;

• The proposed checklist encompasses requirements and recommendations taken from previous proposals, and it further describes quality criteria related to the performance, reliability, reproducibility, and reporting standards of medical ML studies, by also providing relevant references to the literature of interest.

Credit authorship contribution statement

All authors contributed to the conceptualization, drafting of the paper and critical revision.

Declaration of competing interests

The authors have no competing interest.

References


[66] F. Cabitza, A. Campagner, L. M. 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) (2020) 1–21.


