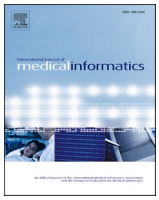
International Journal of Medical Informatics 153 (2021) 104510

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



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

Introducing a comprehensive checklist for the (self)-assessment of medical AI studies

ARTICLE INFO

Keywords
Medical artificial intelligence
Machine learning
Checklist
Quality auditing

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.

As widely known, machine learning (ML) models are beginning to demonstrate early successes in clinical applications [1–3]. Studies that compare the performance of these models and human physicians found that models allegedly perform equally well in many diagnostic and prognostic tasks [4,5]. However, relatively few studies present externally validated results [6–8], and most of them failed to adhere to minimal reporting standards [9,10]. In this respect, poor reporting is one of the main factors preventing studies from being replicated in other settings [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 [12] at an increasing extent, also affects also medical informatics [13], artificial intelligence [14] and its application to medicine [15]. To quote a oft-cited work by Ioannidis [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". 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 4500 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 1400 articles indexed in Scopus; OpenfMRI [20] includes 95 datasets of magnetic resonance imaging (MRI) from more than 3000 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 4000 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 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

[1] http://www.healthdata.org/.

https://doi.org/10.1016/j.ijmedinf.2021.104510
Received 11 May 2021; Received in revised form 26 May 2021; Accepted 27 May 2021
Available online 2 June 2021

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 multiple guidelines, as well as their long production time (as of the writing of this manuscript the TRIPOD-AI extension, which was announced in 2019 [11,47], 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 (see Table 1), 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

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].

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
Problem understanding						
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)? §	○	○	○	○	○	○
2. Is the study design described? (e.g., retrospective, prospective, cross-sectional [51], observational, randomized control trial [52]) §	○	○	○	○	○	○
3. Is the study setting described? (e.g., teaching tertiary hospital; primary care ambulatory, nursing home, medical laboratory, R&D laboratory) §	○	○	○	○	○	○
4. Is the source of data described? (e.g., electronic specialty registry; laboratory information system; electronic health record; picture archiving and communication system) §	○	○	○	○	○	○
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) §	○	○	○	○	○	○
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	○	○	○	○	○	○
Data understanding						
7. Are the subject demographics described in terms of 1. average age (mean or median); 2. age variability (standard deviation (SD) or inter-quartile range (IQR)); 3. gender breakdown (e.g., 55% female, 44% male, 1% not reported); § 4. main comorbidities; 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); 6. socioeconomic status?	○	○	○	○	○	○
8. If the task is supervised, is the gold standard described? (e.g., "100 manually annotated clinical notes and pain scores recorded in EHR, Death, re-admission and International Classification of Disease (ICD) codes in discharge letters"). In particular, the authors should describe the process of ground truthing described in terms of: 1. Number of annotators (raters) producing the labels;	○	○	○	○	○	○

(*continued on next page*)

Table 1 (*continued*)

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
2. Their profession and expertise (e.g., years from specialization or graduation);						
3. Particular instructions given to annotators for quality control (e.g., which data were discarded and why);						
4. Inter-rater agreement score (e.g., Alpha [53], Kappa [54], Rho [17]);						
5. Labelling technique (e.g., majority voting, Delphi method [55], consensus iteration).						
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.						
Data preparation						
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).	○	○	○	○	○	○
Modeling						
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)	○	○	○	○	○	○
Validation						
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 [62];						
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/internal-external model validation procedure described [60,64] (e.g., internal 10-fold CV, time-based cross-validation)? The authors should explicitly specify that the sets have been splitted before normalization, standardization and imputation, to avoid data leakage [61] (also refer to item 17 of this guideline).	○	○	○	○	○	○

(*continued on next page*)

Table 1 (*continued*)

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
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 Ψ [66], Data Representativeness Criterion [68]) 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.						

Table 1 (*continued*)

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
g., true positive fraction, true negative fraction, false positive fraction, false negative fraction).						
5. Reinforcement learning performance, depending on the specific task, should be reported in terms of metrics like [74]: Fixed-Policy Regret, Dispersion across Time, Dispersion across Runs, Risk across Time, Risk across Runs, Dispersion across Fixed-Policy Roll-outs, Risk across Fixed-Policy Rollouts.						
The above estimates should be expressed, whenever possible, with their 95% (or 90%) confidence intervals (CI), or with other indicators of variability [36], 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 (>2×) 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.	○	○	○	○	○	○
Deployment						
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 [77] or similar metrics and present utility curves [78]. 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 [25].	○	○	○	○	○	○
26. Is information regarding model interpretability and explainability available [80] (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 Exlanations Plots	○	○	○	○	○	○

(*continued on next page*)

Table 1 (*continued*)

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
(SHAP) [81]) or model causability [82] should always be supported by some user study, even qualitative or questionnaire-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 (CO_2eq) and details about the estimation method should be given. Any efforts to this end will be appreciated, including those based on tools available online[a], 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[b]	○	○	○	○	○	○
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 [25]; 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 [25]. 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[c] 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].						

[a] https://mlco2.github.io/impact/.

[b] https://www.epa.gov/energy/greenhouse-gas-equivalencies-calculator.

[c] Technology readiness levels (TRL) – Extract from Part 19 – Commission Decision C (2014) 4995.

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).

1. The IJMEDI checklist for assessment of medical AI

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

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.

Conflict of interest statement

The authors declare that there is no conflict of interest regarding the publication of this article.

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.

References

[1] R.C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.
[2] A.L. Fogel, J.C. Kvedar, Artificial intelligence powers digital medicine, NPJ Digit. Med. 1 (1) (2018) 1–4.
[3] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358.
[4] J. Lee, Is artificial intelligence better than human clinicians in predicting patient outcomes? J. Med. Internet Res. 22 (8) (2020) e19918.
[5] J. Shen, C.J. Zhang, B. Jiang, J. Chen, J. Song, Z. Liu, Z. He, S.Y. Wong, P.-H. Fang, W.-K. Ming, Artificial intelligence versus clinicians in disease diagnosis: systematic review, JMIR Med. Inform. 7 (3) (2019) e10010.
[6] R.C. Deo, B.K. Nallamothu, Learning About Machine Learning: The Promise and Pitfalls of Big Data and the Electronic Health Record, 2016.
[7] X. Liu, L. Faes, A.U. Kale, S.K. Wagner, D.J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, et al., A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis, Lancet Digit. Health 1 (6) (2019) e271–e297.
[8] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A.I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, et al., Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans, Nat. Mach. Intell. 3 (3) (2021) 199–217.
[9] R. Aggarwal, V. Sounderajah, G. Martin, D.S. Ting, A. Karthikesalingam, D. King, H. Ashrafian, A. Darzi, Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis, NPJ Digit. Med. 4 (1) (2021) 1–23.
[10] P. Wicks, X. Liu, A.K. Denniston, Going on up to the spirit in AI: will new reporting guidelines for clinical trials of AI interventions improve their rigour? BMC Med. 18 (1) (2020) 1–3.
[11] G.S. Collins, K.G. Moons, Reporting of artificial intelligence prediction models, Lancet 393 (10181) (2019) 1577–1579.
[12] A. Stupple, D. Singerman, L.A. Celi, The reproducibility crisis in the age of digital medicine, NPJ Digit. Med. 2 (1) (2019) 1–3.
[13] E. Coiera, E. Ammenwerth, A. Georgiou, F. Magrabi, Does health informatics have a replication crisis? J. Am. Med. Inform. Assoc. 25 (8) (2018) 963–968.
[14] M. Hutson, Artificial intelligence faces reproducibility crisis, Science 359 (6377) (2018) 725–726, https://doi.org/10.1126/science.359.6377.725.
[15] A.L. Beam, A.K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care, JAMA 323 (4) (2020) 305–306.
[16] J.P. Ioannidis, Why most published research findings are false, PLoS Med. 2 (8) (2005) e124.
[17] F. Cabitza, A. Campagner, D. Albano, A. Aliprandi, A. Bruno, V. Chianca, A. Corazza, F. Di Pietto, A. Gambino, S. Gitto, et al., The elephant in the machine: proposing a new metric of data reliability and its application to a medical case to assess classification reliability, Appl. Sci. 10 (11) (2020) 4014.
[18] M.A. Gianfrancesco, S. Tamang, J. Yazdany, G. Schmajuk, Potential biases in machine learning algorithms using electronic health record data, JAMA Intern. Med. 178 (11) (2018) 1544–1547.
[19] A.E. Johnson, T.J. Pollard, L. Shen, H.L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L.A. Celi, R.G. Mark, Mimic-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9.
[20] R.A. Poldrack, D.M. Barch, J. Mitchell, T. Wager, A.D. Wagner, J.T. Devlin, C. Cumba, O. Koyejo, M. Milham, Toward open sharing of task-based fmri data: the openfmri project, Front. Neuroinform. 7 (2013) 12.
[21] K. Yan, X. Wang, L. Lu, R.M. Summers, Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning, J. Med. Imaging 5 (3) (2018) 036501.
[22] C.H. Lee, H.-J. Yoon, Medical big data: promise and challenges, Kidney Res. Clin. Pract. 36 (1) (2017) 3.
[23] J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, L.A. Celi, The myth of generalisability in clinical research and machine learning in health care, Lancet Digit. Health 2 (9) (2020) e489–e492.
[24] E. Christodoulou, J. Ma, G.S. Collins, E.W. Steyerberg, J.Y. Verbakel, B. Van Calster, A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models, J. Clin. Epidemiol. 110 (2019) 12–22.
[25] I. Scott, S. Carter, E. Coiera, Clinician checklist for assessing suitability of machine learning applications in healthcare, BMJ Health Care Inform. 28 (1) (2021).
[26] T. Ching, D.S. Himmelstein, B.K. Beaulieu-Jones, A.A. Kalinin, B.T. Do, G.P. Way, E. Ferrero, P.-M. Agapow, M. Zietz, M.M. Hoffman, et al., Opportunities and obstacles for deep learning in biology and medicine, J. R. Soc. Interface 15 (141) (2018) 20170387.
[27] C.J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, D. King, Key challenges for delivering clinical impact with artificial intelligence, BMC Med. 17 (1) (2019) 1–9.
[28] F. Cabitza, R. Rasoini, G.F. Gensini, Unintended consequences of machine learning in medicine, JAMA 318 (6) (2017) 517–518.
[29] W.G. Miller, Harmonization: its time has come, Clin. Chem. 63 (7) (2017).
[30] A. Coskun, F. Braga, A. Carobene, X.T. Ganduxe, A.K. Aarsand, P. Fernández-Calle, J.D.-G. Marco, W. Bartlett, N. Jonker, B. Aslan, et al., Systematic review and meta-analysis of within-subject and between-subject biological variation estimates of 20 haematological parameters, Clin. Chem. Lab. Med. (CCLM) 58 (1) (2020) 25–32.
[31] T.-M. Chan, Y. Li, C.-C. Chiau, J. Zhu, J. Jiang, Y. Huo, Imbalanced target prediction with pattern discovery on clinical data repositories, BMC Med. Inform. Decis. Mak. 17 (1) (2017) 1–12.
[32] A. Vabalas, E. Gowen, E. Poliakoff, A.J. Casson, Machine learning algorithm validation with a limited sample size, PLOS ONE 14 (11) (2019) e0224365.
[33] F. Cismondi, A.S. Fialho, S.M. Vieira, S.R. Reti, J.M. Sousa, S.N. Finkelstein, Missing data in medical databases: impute, delete or classify? Artif. Intell. Med. 58 (1) (2013) 63–72.
[34] C.F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J.R.G. Marquéz, B. Gruber, B. Lafourcade, P.J. Leitao, et al., Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography 36 (1) (2013) 27–46.
[35] A. Subbaswamy, S. Saria, From development to deployment: dataset shift, causality, and shift-stable models in health AI, Biostatistics 21 (2) (2020) 345–352.
[36] X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, et al., Accounting for variance in machine learning benchmarks, Proc. Mach. Learn. Syst. 3 (2021).
[37] J. Waring, C. Lindvall, R. Umeton, Automated machine learning: review of the state-of-the-art and opportunities for healthcare, Artif. Intell. Med. 104 (2020) 101822.
[38] S. Vollmer, B.A. Mateen, G. Bohner, F.J. Király, R. Ghani, P. Jonsson, S. Cumbers, A. Jonas, K.S. McAllister, P. Myles, et al., Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, BMJ 368 (2020).
[39] H. Ibrahim, X. Liu, A.K. Denniston, Reporting guidelines for artificial intelligence in healthcare research, Clin. Exp. Ophthalmol. (2021).
[40] S.C. Rivera, X. Liu, A.-W. Chan, A.K. Denniston, M.J. Calvert, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-AI extension, BMJ 370 (2020).
[41] X. Liu, S.C. Rivera, D. Moher, M.J. Calvert, A.K. Denniston, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-AI extension, BMJ 370 (2020).
[42] B. Norgeot, G. Quer, B.K. Beaulieu-Jones, A. Torkamani, R. Dias, M. Gianfrancesco, R. Arnaout, I.S. Kohane, S. Saria, E. Topol, et al., Minimum information about clinical artificial intelligence modeling: the mi-claim checklist, Nat. Med. 26 (9) (2020) 1320–1324.
[43] C. Johner, P. Balachandran, L. Oala, A.Y. Lee, A. Leite, M. Werneck, L.A. Andrew, C. Molnar, J. Rumball-Smith, P. Baird, P.G. Goldschmidt, P. Quartarolo, S. Xu, S. Piechottka, Z. Hornberger, Good practices for health applications of machine learning: considerations for manufacturers and regulators, in: L. Oala (Ed.), ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) – Meeting K, Vol. K, ITU, 2021.
[44] L. Oala, J. Fehr, L. Gilli, P. Balachandran, A.W. Leite, S. Calderon-Ramirez, D.X. Li, G. Nobis, E.A.M. Alvarado, G. Jaramillo-Gutierrez, et al., Ml4h auditing: from paper to practice. Machine Learning for Health, PMLR, 2020, pp. 280–317.
[45] R.F. Wolff, K.G. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, Probast: a tool to assess the risk of bias and applicability of prediction model studies, Ann. Intern. Med. 170 (1) (2019) 51–58.
[46] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement, Circulation 131 (2) (2015) 211–219.
[47] X. Liu, L. Faes, M.J. Calvert, A.K. Denniston, Extension of the consort and spirit statements, Lancet 394 (10205) (2019) 1225.
[48] V. Sounderajah, H. Ashrafian, R. Aggarwal, J. De Fauw, A.K. Denniston, F. Greaves, A. Karthikesalingam, D. King, X. Liu, S.R. Markar, et al., Developing specific reporting guidelines for diagnostic accuracy studies assessing ai interventions: the stard-AI steering group, Nat. Med. 26 (6) (2020) 807–808.
[49] T. Hernandez-Boussard, S. Bozkurt, J.P. Ioannidis, N.H. Shah, Minimar (minimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care, J. Am. Med. Inform. Assoc. 27 (12) (2020) 2011–2015.
[50] R. Wirth, J. Hipp, Crisp-dm: towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, vol. 1, Springer-Verlag, London, UK, 2000.
[51] J.I. Hudson, H.G. Pope Jr., R.J. Glynn, The cross-sectional cohort study: an underutilized design, Epidemiology 16 (3) (2005) 355–359.
[52] E.L. Hannan, Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations, JACC Cardiovasc. Interv. 1 (3) (2008) 211–217.
[53] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, Sage Publications, 2018.
[54] J.L. Fleiss, Measuring nominal scale agreement among many raters, Psychol. Bull. 76 (5) (1971) 378.
[55] H.A. Linstone, M. Turoff, et al., The Delphi Method, Addison-Wesley Reading, MA, 1975.
[56] C.J. McDonald, S.M. Huff, J.G. Suico, G. Hill, D. Leavelle, R. Aller, A. Forrey, K. Mercer, G. DeMoor, J. Hook, et al., Loinc, a universal standard for identifying laboratory observations: a 5-year update, Clin. Chem. 49 (4) (2003) 624–633.
[57] R.-D. Treede, W. Rief, A. Barke, Q. Aziz, M.I. Bennett, R. Benoliel, M. Cohen, S. Evers, N.B. Finnerup, M.B. First, et al., A classification of chronic pain for icd-11, Pain 156 (6) (2015) 1003.
[58] R. Cornet, N. de Keizer, Forty years of snomed: a literature review, BMC Med. Inform. Decis. Mak. 8 (1) (2008) 1–6.

[59] A.K. Waljee, A. Mukherjee, A.G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu, P.D. Higgins, Comparison of imputation methods for missing laboratory data in medicine, BMJ Open 3 (8) (2013).
[60] S. Borra, A. Di Ciaccio, Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods, Comput. Stat. Data Anal. 54 (12) (2010) 2976–2989.
[61] S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: formulation, detection, and avoidance, ACM Trans. Knowl. Discov. Data (TKDD) 6 (4) (2012) 1–21.
[62] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, H. Larochelle, Improving Reproducibility in Machine Learning Research (a Report From the Neurips 2019 Reproducibility Program), 2020 (arXiv preprint), arXiv:2003.12206.
[63] B. Van Calster, D.J. McLernon, M. Van Smeden, L. Wynants, E.W. Steyerberg, Calibration: the achilles heel of predictive analytics, BMC Med. 17 (1) (2019) 1–7.
[64] E.W. Steyerberg, F.E. Harrell Jr., Prediction models need appropriate internal, internal-external, and external validation, J. Clin. Epidemiol. 69 (2016) 245.
[65] I. Balki, A. Amirabadi, J. Levman, A.L. Martel, Z. Emersic, B. Meden, A. Garcia-Pedrero, S.C. Ramirez, D. Kong, A.R. Moody, et al., Sample-size determination methodologies for machine learning in medical imaging research: a systematic review, Can. Assoc. Radiol. J. 70 (4) (2019) 344–353.
[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 Med. Inform. Decis. Mak. 20 (1) (2020) 1–21.
[67] S. Bleeker, H. Moll, E. Steyerberg, A. Donders, G. Derksen-Lubsen, D. Grobbee, K. Moons, External validation is necessary in prediction research: a clinical example, J. Clin. Epidemiol. 56 (9) (2003) 826–832.
[68] E. Schat, R. van de Schoot, W.M. Kouw, D. Veen, A.M. Mendrik, The data representativeness criterion: predicting the performance of supervised classification based on data set similarity, PLOS ONE 15 (8) (2020) e0237009.
[69] K.I. Snell, L. Archer, J. Ensor, L.J. Bonnett, T.P. Debray, B. Phillips, G.S. Collins, R. D. Riley, External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb, J. Clin. Epidemiol. 135 (2021) 79–89.
[70] B. Ozenne, F. Subtil, D. Maucort-Boulch, The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases, J. Clin. Epidemiol. 68 (8) (2015) 855–859.
[71] D. Chicco, N. Tötsch, G. Jurman, The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation, BioData Min. 14 (1) (2021) 1–22.
[72] E. Rendón, I. Abundez, A. Arizmendi, E.M. Quiroz, Internal versus external cluster validation indexes, Int. J. Comput. Commun. 5 (1) (2011) 27–34.
[73] A. Fenster, B. Chiu, Evaluation of segmentation algorithms for medical imaging, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, IEEE, 2006, pp. 7186–7189.
[74] S.C. Chan, S. Fishman, J. Canny, A. Korattikara, S. Guadarrama, Measuring the reliability of reinforcement learning algorithms, in: International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020.
[75] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[76] K.E. Taylor, Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res. Atmos. 106 (D7) (2001) 7183–7192.
[77] A.J. Vickers, B. Van Calster, E.W. Steyerberg, Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests, BMJ 352 (2016).
[78] B. Van Calster, L. Wynants, J.F. Verbeek, J.Y. Verbakel, E. Christodoulou, A. J. Vickers, M.J. Roobol, E.W. Steyerberg, Reporting and interpreting decision curve analysis: a guide for investigators, Eur. Urol. 74 (6) (2018) 796–804.
[79] L. Wynants, M. van Smeden, D.J. McLernon, D. Timmerman, E.W. Steyerberg, B. Van Calster, Three myths about risk thresholds for prediction models, BMC Med. 17 (1) (2019) 1–7.
[80] A. Vellido, The importance of interpretability and visualization in machine learning for applications in medicine and health care, Neural Comput. Appl. (2019) 1–15.
[81] M. Sundararajan, A. Najmi, The many Shapley values for model explanation, International Conference on Machine Learning (2020) 9269–9278.
[82] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, H. Müller, Causability and explainability of artificial intelligence in medicine, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 9 (4) (2019) e1312.
[83] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs), KI-Künstliche Intell. (2020) 1–6.
[84] R.R. Hoffman, S.T. Mueller, G. Klein, J. Litman, Metrics for Explainable AI: Challenges and Prospects, 2018 (arXiv preprint), arXiv:1812.04608.
[85] C. Schwartzenberg, T. van Engers, Y. Li, The fidelity of global surrogates in interpretable machine learning, BNAIC/BeneLearn 2020 (2020) 269.
[86] E. Vayena, A. Blasimme, I.G. Cohen, Machine learning in medicine: addressing ethical challenges, PLOS Med. 15 (11) (2018) e1002689.
[87] A. Rajkomar, M. Hardt, M.D. Howell, G. Corrado, M.H. Chin, Ensuring fairness in machine learning to advance health equity, Ann. Intern. Med. 169 (12) (2018) 866–872.
[88] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, C. Ré, Hidden stratification causes clinically meaningful failures in machine learning for medical imaging, Proceedings of the ACM Conference on Health, Inference, and Learning (2020) 151–159.
[89] J. Cowls, A. Tsamados, M. Taddeo, L. Floridi, The AI Gambit-Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and Recommendations, 2021. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3804983.
[90] B. Van Calster, L. Wynants, D. Timmerman, E.W. Steyerberg, G.S. Collins, Predictive analytics in health care: how can we know it works? J. Am. Med. Inform. Assoc. 26 (12) (2019) 1651–1654.
[91] D.G. Fryback, J.R. Thornbury, The efficacy of diagnostic imaging, Med. Decis. Mak. 11 (2) (1991) 88–94.
[92] K.G. van Leeuwen, S. Schalekamp, M.J. Rutten, B. van Ginneken, M. de Rooij, Artificial intelligence in radiology: 100 commercially available products and their scientific evidence, Eur. Radiol. (2021) 1–8.
[93] A. Lavin, C.M. Gilligan-Lee, A. Visnjic, S. Ganju, D. Newman, S. Ganguly, D. Lange, A.G. Baydin, A. Sharma, A. Gibson, et al., Technology Readiness Levels for Machine Learning Systems, 2021 (arXiv preprint), arXiv:2101.03989.
[94] S.E. Davis, R.A. Greevy, T.A. Lasko, C.G. Walsh, M.E. Matheny, Comparison of prediction model performance updating protocols: using a data-driven testing procedure to guide updating, AMIA Annual Symposium Proceedings, Vol. 2019 (2019) 1002.

Federico Cabitza*, Andrea Campagner
DISCo, University of Milano-Bicocca, viale Sarca 336, Milano 20126, Italy

* Corresponding author.
E-mail address: federico.cabitza@unimib.it (F. Cabitza).