The IJMI medical machine learning checklist:

0: absent; 1: inadequately addressed; 2: sufficiently addressed; 3: adequately addressed.

	Requirement	0	1	2	3	NA
1.	**Is the study population described?** (e.g., patients admitted at emergency department; all patients) §	○	○	○	√	○
2.	**Are the inclusion / exclusion criteria described** (e.g., patients older than 18 tested for COVID-19; all inpatients hospitalized for 24 or more hours) §	○	○	○	√	○
3.	**Is the study setting described?** (e.g., teaching tertiary hospital; primary care ambulatory, nursing home, medical laboratory, R&D lab) §	○	○	○	√	○
4.	**Is the source of data described?** (e.g., electronic speciality registry; laboratory information system, Electronic Health Record, Picture Archiving and Communication system) § Any consideration about the data quality of the source (e.g., completeness, plausibility, robustness with respect to upcoding or downcoding practices) is advocated promoted and appreciated.	○	○	○	√	○
5.	**Is the subject demographics described** in terms of a. average age (mean or median), b. age variability (standard deviation or IQR) c. gender breakdown (e.g., 55% female, 44% male, 1% not reported)? §	○	○	○	√	○
6.	**Is the subject demographics described** in further details, like main comorbidities, race (e.g., American Indian or Alaska Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Other Pacific Islander, White), ethnicity (e.g., European, or North African) and Socioeconomic status? §	○	○	√	○	○
7.	**Is the model task reported?** (e.g., Binary Classification, multi-class classification, multi-label classification, ordinal regression, continuous regression, clustering, dimensionality reduction, segmentation) §	○	○	○	√	○
8.	**Is the medical task reported?** (e.g., diagnostic detection, diagnostic characterization, diagnostic stadiation, prognosis -on what endpoint-, treatment planning, monitoring) §	○	○	○	√	○
9.	**Is the model output specified?** (e.g., COVID-19 positivity probability score; probability of infection within 5 days; Postoperative 3-month pain scores) §	○	○	○	√	○
10.	**Is the target user indicated?** (e.g., clinician, radiologist, hospital management	○	○	○	√	○

	team, insurance company, patients) §					
11.	**Is the data splitting described** (no data splitting, k-fold cross-validation, Nested k-fold CV, Repeated cross-validation, Bootstrap Validation, Leave-one-out CV, 80%/10%10% train/validation/test)? In case of data splitting, the authors should explicitly state that splitting was performed before any pre-processing (e.g. normalization, standardization, missing value imputation, feature selection) or model construction (training, hyper-parameter optimization) steps, in order to avoid data leakage[1] and overfitting.	○	○	○	√	○
12.	If supervised, **is the gold standard described?** (e.g., "100 manually annotated clinical notes and pain scores recorded in EHR, Death, re-admission and ICD codes in EHRs") §	○	○	○	√	○
13.	**Is the process of ground truthing described** in terms of a. Number of annotators (raters) producing the labels b. their profession and expertise (e.g., years from specialization or graduation) c. particular instructions given to annotators for quality control (e.g., what data were discarded and why) d. inter-rater agreement score (e.g., Alpha[2], Kappa[3], Rho[4]) e. labelling technique (e.g., majority voting, Delphi method, consensus iteration)	○	○	○	○	√
14.	**Is the model architecture or type described?** (e.g., SVM, Random Forest, Boosting, Logistic Regression, Nearest Neighbors, Convolutional Neural Network)	○	○	○	√	○
15.	In 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 should be done for all or, in case these are more than 20, for a significant subset of the most predictive features in the following terms: name, short description, type (nominal, ordinal, continuous), and a. if continuous: unit of measure, range (min, max), mean and standard deviation (or median and	○	○	√	○	○

[1] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *6*(4), 1-21.

[2] Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

[3] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, *76*(5), 378.

[4] Cabitza, F., Campagner, A., Albano, D., Aliprandi, A., Bruno, A., Chianca, V., ... & Sconfienza, L. M. (2020). The elephant in the machine: Proposing a new metric of data reliability and its application to a medical case to assess classification reliability. *Applied Sciences*, *10*(11), 4014.

	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. b. If nominal, all codes/values and their distribution. Feature transformation (e.g. one-hot encoding) should be reported if applied. Any terminology standard should explicitly be mentioned (e.g., LOINC, ICD-11, SNOMED) if applied.					
16.	**Is outlier detection and analysis performed and reported?** If this is the case, the definition of outlier should be given and the techniques applied to manage them should be described (e.g., removal through application of an Isolation Forest model).	○	○	○	○	√
17.	**Is missing-value management described?** This should be done in the following terms: a. the missing rate for each feature should be reported. b. the technique of imputation, if any, should be described, and reasons given for its choice (e.g., missing data were imputed using median of the variable distribution). If the missing rate is higher than 10%, a reflection about the impact on performance of a technique with respect to others would be appreciable[5] c. If records have been deleted for their low completeness, the similarity of these cases with respect to the remaining sample should be assessed.	○	○	○	√	○
18.	**Is the model training and selection described?** In particular, the hyper-parameter or model selection should be described in terms of a. range of hyper-parameters[6], b. method used to select the best hyper-parameter configuration (e.g., Hyper-parameter selection was performed through nested k-fold CV based grid search), c. full specification of the hyper-parameters used to generate results[25].	○	○	○	√	○

[5] Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., ... & Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, *3*(8).

[6] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d'Alché-Buc, F., ... & Larochelle, H. (2020). Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *arXiv preprint arXiv:2003.12206*.

	d. the procedure (if any) to limit over-fitting.					
19.	(for classification models) **is the model calibration described?** In this case, the Brier score should be reported, and a calibration plot should be presented[7].	○	○	○	√	○
20.	(For classification models), **is the utility of the model discussed?** To this aim, the authors should report the performance of a baseline model (e.g., logistic regression, Naive Bayes) or recall the performance of a random classifier. Additionally, the authors could report the Net Benefit[8] or similar metrics and present utility curves[9]. The authors should be encouraged to discuss the selection of appropriate risk thresholds; the relative value of benefits (true positives/negatives) and harms (false positives/negatives); and the clinical utility of the proposed models.	○	○	○	√	○
21.	**Is the internal/internal-external model validation procedure described** (e.g., Internal 10-fold cross-validation, random Hold-out validation set, Time-based cross-validation)? The authors should explicitly specify that the sets have been splitted before normalization, standardization and imputation, to avoid data leakage[20] (see also item 11 of this guideline). Moreover, the authors should try to choose the test set so that it is the most diverse with respect to the rest of the sample (w.r.t. some multivariate similarity function) and how this choice relates with conservative (and lower-bound) estimates of the model's accuracy (and performance). If performance on external datasets is found to be similar (or even better) than on training and internal datasets, the authors should provide some explanatory conjectures why this happened (e.g., high heterogeneity of the training set, high homogeneity or the external dataset).	○	○	○	√	○
22.	**Has been the model externally validated?** In this case, the characteristics of the external validation set(s) should be described. For instance, the authors could comment about the heterogeneity of the data wrt the training set (e.g., degree of	○	○	○	√	○

[7] Van Calster, B., & Vickers, A. J. (2015). Calibration of risk prediction models: impact on decision-analytic performance. *Medical decision making*, *35*(2), 162-169.

[8] Vickers, A. J., Van Calster, B., & Steyerberg, E. W. (2016). Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *bmj*, *352*.

[9] Van Calster, B., Wynants, L., Verbeek, J. F., Verbakel, J. Y., Christodoulou, E., Vickers, A. J., ... & Steyerberg, E. W. (2018). Reporting and interpreting decision curve analysis: a guide for investigators. *European urology*, *74*(6), 796-804.

correspondence Ψ[10], Data Representativeness Criterion[11]) and about the cardinality of the external sample[12].					
23. **Are the main error-based metrics used?** a. Classification performance must be reported in terms of i. Accuracy, ii. Balanced accuracy; iii. Specificity; iv. Sensitivity; v. 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" instead or in addition to AUROC, that is the area under the 'Positive Predictive Value' and 'Sensitivity' curve[13]) vi. Optionally: F1 score, Matthew coefficient[14], F score of sensitivity and specificity, the full confusion matrix. b. Regression performance should be reported in terms of R^2, MAE, RMSE; ratio between MAE/RMSE and SD (of the target). c. Clustering performance should be reported in terms of: i. External validation metrics (when ground truth labels are available): e.g. mutual information, purity, Rand index. ii. Internal validation metrics (e.g. Davies-Bouldin index, Silhouette index, Homogeneity): since internal validation metrics are usually algorithm-	○	○	○	√	○

[10] Cabitza, F., Campagner, A., & Sconfienza, L. M. (2020). As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. *BMC Medical Informatics and Decision Making*, *20*(1), 1-21.

[11] Schat, E., van de Schoot, R., Kouw, W. M., Veen, D., & Mendrik, A. M. (2020). The data representativeness criterion: Predicting the performance of supervised classification based on data set similarity. *Plos one*, *15*(8), e0237009.

[12] Snell, K. I., Archer, L., Ensor, J., Bonnett, L. J., Debray, T. P., Phillips, B., ... & Riley, R. D. (2021). External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *Journal of clinical epidemiology*, *135*, 79-89.

[13] Ozenne, B., Subtil, F., & Maucort-Boulch, D. (2015). The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases. *Journal of clinical epidemiology*, *68*(8), 855-859.

[14] Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, *14*(1), 1-22.

dependent, the reported results should be discussed d. The above estimates for points a, b and c should be expressed, whenever possible, with their 95% (or 90%) confidence intervals, or with other indicators of variability, 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 cross-validation or bootstrap was applied, normal approximation).					
24. **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 the cases represent statistical outliers for some covariate, the authors should comment on that.	○	○	○	○	√
25. **Is information regarding model interpretability available**[15] (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 SHAP (SHapley Additive exPlanations plots) or model causability[16] should always be supported by some user study, even qualitative or questionnaire-based.	○	○	√	○	○
26. **Is there any discussion regarding model fairness, ethical concerns or risks of bias**[17,18] (for a list of clinically relevant biases see [19])? If possible, the authors should report the model performance stratified for particularly relevant population strata (e.g. model performance on male vs female subjects, (e.g. model performance on male vs female subjects, or on minority groups).	○	○	○	○	√
27. **Is any point made about the environmental sustainability of the**	○	○	○	○	√

[15] Vellido, A. (2019). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural computing and applications*, 1-15.

[16] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *9*(4), e1312.

[17] Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, *15*(11), e1002689.

[18] Scott, I., Carter, S., & Coiera, E. (2021). Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health & Care Informatics*, *28*(1).

[19] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872

	model or about the carbon footprint[20] of either the training phase or inference phase (use) of the model? If this is the case, such a footprint should be expressed in terms of carbon dioxide equivalent (CO2eq) and details about the estimation method should be given. To this aim, any efforts should be appropriately appreciated and promoted, including those based on tools available online[21], 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[22].					
28.	**Is code and data shared with the community?** § If not, are reasons given? If code and data are shared, institutional repositories such as Zenodo should be preferred to private-owned ones (arxiv, GitHub). If code is shared, specification of dependencies should be reported[25] and a clear distinction between training code and evaluation code should be made[25].	√	○	○	○	○
29.	**Is either a sand-box or a fully-operating system made freely accessible on the Web to test the system?**	○	○	○	√	○
30.	**Is the system already adopted in daily practice?** If this is the case, where (setting name) and since when. Also a qualitative assessment of the *level of efficacy* of the the contribution of the AI software to the clinical process would be appreciated, e.g., by referring to a model like the one proposed in[23] and recently adapted in[24]. If this is not the case, an assessment of the technology readiness of the described system should be proposed, with explicit reference to the Technology Readiness Level (TRL[25]) framework or to any adaptation of this framework to the AI/ML domain[26].	○	○	○	○	√

§ inspired by the MINIMAR guidelines, Hernandez-Boussard, T., Bozkurt, S., Ioannidis, J. P., & Shah, N. H. (2020). MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care. Journal of the American Medical Informatics Association, 27(12), 2011-2015.

[20] Cowls, J., Tsamados, A., Taddeo, M., & Floridi, L. (2021). The AI Gambit—Leveraging Artificial Intelligence to Combat Climate Change: Opportunities, Challenges, and Recommendations. *Challenges, and Recommendations (March 15, 2021)*.

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

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

[23] Fryback DG, Thornbury JR (1991) The efficacy of diagnostic imaging. Med Decis Making 11:88–94

[24] van Leeuwen, K. G., Schalekamp, S., Rutten, M. J., van Ginneken, B., & de Rooij, M. (2021). Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. European Radiology, 1-8.

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

[26] Lavin, A., Gilligan-Lee, C. M., Visnjic, A., Ganju, S., Newman, D., Ganguly, S., ... & Parr, J. (2021). Technology readiness levels for machine learning systems. *arXiv preprint arXiv:2101.03989*.