ChAMAI checklist – Checklist for assessment of medical AI [1]

Checklist for assessment of requirements and recommendations for sound Machine Learning contributions to the existing literature, with a focus on medical applications.
Items in bold indicate priority aspects to be considered. Items denoted with a § symbol are directly inspired by the MINIMAR guideline [2]. The section names for the checklist items are directly inspired by the CRISP-DM framework [3].

For use as an article-review support tool NA: not applicable; OK: adequately addressed; mR: minor revisions needed; MR: major revisions needed.

For use as a literature systematic review support tool Please assign 2, 1 e 0 points for high-priority requirements that are, respectively, OK, mR and MR; assign half these scores (i.e., 1, 0.5 and 0) for low-priority requirements. Points can be added up for section-level scores and total scores for comparative (both cross-article and longitudinal) purposes.

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 [4], observational, randomized control trial [5]) §			○	○	○	○
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	○	○	○	○	○	○

[1] Including some NLP tasks, such as named entity recognition, anonymization and text classification [1]

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
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?** **N.B. in the NLP case, subject demographics could be related to the text producers (if applicable) and it could encompass the source context of the unstructured data. It is important to specify the application domain and if it is a language dependent task.**			○	○	○	○
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;** **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 [6], Kappa [7], Rho [8]);** **5. Labelling technique (e.g., majority voting, Delphi method [9], 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 [10], ICD-11 [11], SNOMED [12]) if applied.**	○		○	○	○	○
Data Preparation						
10. If performed, is outlier detection reported? If the answer is yes, the definition of an outlier should be given [13] and the techniques applied to manage outliers should be described (e.g., removal through the application of an Isolation Forest model, or for NLP applications, of an excessively long/short text).	○		○	○	○	○

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
11. If applicable, 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 [14].**	○		○	○	○	○
12. If performed, is feature pre-processing 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). While for NLP task (if needed): stemming, lemmatization, stop words removal, tokenization, etc.. It is appropriate to describe the length of the input vector, specifically the number of tokens used. [15]	○		○	○	○	○
13. If applicable, 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, balanced batch).	○		○	○	○	○
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, terms of clinical terms to be identified) §			○	○	○	○
16. Is the model architecture or type described? (e.g., SVM, Random Forest, Boosting, Logistic Regression, Nearest Neighbors, Convolutional/Recurrent Neural Network, K-Means, Generative Adversarial Network, Bayesian Network, Transformer, Latent Dirichlet Allocation)			○	○	○	○
Validation						
17. Is the data splitting [16] 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 [13])? 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, sampling) or model construction steps (training, hyper-parameter optimization), so to avoid data leakage [17] 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 [18];** **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 [18];** **4. Procedure (if any) to limit over-fitting, in particular as related to the sample size [19].**			○	○	○	○

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
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 [20]	○		○	○	○	○
20. Is the internal/internal-external model validation procedure described [16, 21] (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 [17] (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 [22, 19]. 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 [23] (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[24]).	○		○	○	○	○
21. Has the model been externally validated [25]? If the answer is yes, the characteristics of the external validation set(s) should be described [26]. For instance, the authors could comment about the heterogeneity of the data with respect to the training set (e.g., degree of correspondence Ψ [23], Data Representativeness Criterion [27]) and the cardinality of the external sample [28]. 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)	○	○	○	○	○	○

Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
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" [29]). Optionally also in terms of: positive and negative predictive value, F1 score, Matthew coefficient [30, 31], F score of sensitivity and specificity, the full confusion matrix, Hamming Loss (for multi-label classification), Jaccard Index (for multi-label classification). For classification tasks, a rough guideline to qualitatively evaluate classification scores is proposed in [24].** **2. Regression performance should be reported in terms of: R^2 [32]; Mean Absolute Error (MAE); Root Mean Square Error (RMSE); 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, Topic Coherence). The reported results of internal validation metrics should be discussed [33]** **4. Image segmentation performance, depending on the specific task, should be reported in terms of metrics like [34]: 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).** **5. Reinforcement learning performance, depending on the specific task, should be reported in terms of metrics like [35]: Fixed-Policy Regret, Dispersion across Time, Dispersion across Runs, Risk across Time, Risk across Runs, Dispersion across Fixed-Policy Rollouts, 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 [37] (e.g. through CI comparisons, or hypothesis testing). When comparing multiple regression models, a Taylor diagram [38] could be reported and discussed.**			○	○	○	○
23. Are some relevant errors described? The authors should describe the characteristic of some noteworthy classification errors [39] 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.	○	○	○	○	○	○
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 [40] or similar metrics and present utility curves [41]. In particular, the authors are encouraged to discuss the selection of appropriate risk thresholds [42]; the relative value of benefits (true positives/negatives) and harms (false positives/negatives); and the clinical utility of the proposed models [19].	○	○	○	○	○	○

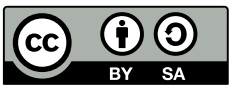
Requirement	Authors			Reviewers		
	NA	No	Yes	OK	mR	MR
26. Is information regarding model interpretability and explainability available [43] (e.g. feature importance, interpretable surrogate models, information about the model parameters)? Claims of "high" or "adequate" model interpretability (e.g., by means of visual aids like decision trees, Variable Importance Plots or Shapley Additive Exlanations Plots (SHAP) [44], Attention scores from a Transformer architecture [45]) or model causability [46] should always be supported by some user study, even qualitative or questionnaire-based [47]. In the case surrogate models were applied, the authors should report about their fidelity [48, 49]. .	○	○	○	○	○	○
27. Is there any discussion regarding model fairness, ethical concerns or risks of bias [19, 50] (for a list of clinically relevant biases, refer to [51])? If possible, the authors should report the model performance stratified for particularly relevant population strata [52] (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 [53], 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[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 software code and data shared with the community [18, 54, 55]? § If not, are reasons given? If software code and data are shared, institutional repositories such as Zenodo should be preferred to private-owned repositories (FigShare and arxiv for the datasets, GitHub, GitLab, or SourceForge for the code). If software code is shared, specification of dependencies should be reported and a clear distinction between training code and evaluation code should be made [56]. 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. as a side note, open source programming languages, such as Python or R, should be preferred over proprietary ones.			○	○	○	○
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 [19]; a qualitative assessment of the level of efficacy of the system's contribution to the clinical process (e.g., [57, 58]); any comment about the technical and staff training effort actually required [19]. 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 [59]. In either above cases (yes/no), the authors should report about the procedures (if any) for performance monitoring, model maintenance and updating [60].	○	○	○	○	○	○

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

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

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

If you want to cite this checklist: Cabitza F., Campagner, A. (2021) The need to separate the wheat from the chaff in medical informatics. International Journal of Medical Informatics.



References

[1] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, H. Xu, Deep learning in clinical natural language processing: a methodical review, Journal of the American Medical Informatics Association 27 (3) (2019) 457–470. arXiv:https://academic.oup.com/jamia/article-pdf/27/3/457/34152802/ocz200.pdf, doi:10.1093/jamia/ocz200.
URL https://doi.org/10.1093/jamia/ocz200

[2] 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, Journal of the American Medical Informatics Association 27 (12) (2020) 2011–2015.

[3] R. Wirth, J. Hipp, Crisp-dm: Towards a standard process model for data mining, in: Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, Vol. 1, Springer-Verlag London, UK, 2000.

[4] J. I. Hudson, H. G. Pope Jr, R. J. Glynn, The cross-sectional cohort study: an underutilized design, Epidemiology 16 (3) (2005) 355–359.

[5] E. L. Hannan, Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations, JACC: Cardiovascular Interventions 1 (3) (2008) 211–217.

[6] K. Krippendorff, Content analysis: An introduction to its methodology, Sage publications, 2018.

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

[8] 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, Applied Sciences 10 (11) (2020) 4014.

[9] H. A. Linstone, M. Turoff, et al., The delphi method, Addison-Wesley Reading, MA, 1975.

[10] 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, Clinical chemistry 49 (4) (2003) 624–633.

[11] 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.

[12] R. Cornet, N. de Keizer, Forty years of snomed: a literature review, BMC Medical Informatics and Decision Making 8 (1) (2008) 1–6.

[13] D. Chicco, Ten quick tips for machine learning in computational biology, BioData mining 10 (1) (2017) 1–17.

[14] 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).

[15] K. Smelyakov, D. Karachevtsev, D. Kulemza, Y. Samoilenko, O. Patlan, A. Chupryna, Effectiveness of preprocessing algorithms for natural language processing applications, in: 2020 IEEE International Conference on Problems of Infocommunications. Science and Technology (PIC S T), 2020, pp. 187–191. doi:10.1109/PICST51311.2020.9467919.

[16] S. Borra, A. Di Ciaccio, Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods, Computational statistics & data analysis 54 (12) (2010) 2976–2989.

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

[18] 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), arXiv preprint arXiv:2003.12206 (2020).

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

[20] B. Van Calster, D. J. McLernon, M. Van Smeden, L. Wynants, E. W. Steyerberg, Calibration: the achilles heel of predictive analytics, BMC medicine 17 (1) (2019) 1–7.

[21] E. W. Steyerberg, F. E. Harrell Jr, Prediction models need appropriate internal, internal-external, and external validation, Journal of clinical epidemiology 69 (2016) 245.

[22] 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, Canadian Association of Radiologists Journal 70 (4) (2019) 344–353.

[23] 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.

[24] F. Cabitza, A. Campagner, F. Soares, L. G. de Guadiana-Romualdo, F. Challa, A. Sulejmani, M. Seghezzi, A. Carobene, The importance of being external. methodological insights for the external validation of machine learning models in medicine, Computer Methods and Programs in Biomedicine 208 (2021) 106288.

[25] 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, Journal of clinical epidemiology 56 (9) (2003) 826–832.

[26] I. Walsh, D. Fishman, D. Garcia-Gasulla, T. Titma, G. Pollastri, J. Harrow, F. E. Psomopoulos, S. C. Tosatto, Dome: recommendations for supervised machine learning validation in biology, Nature methods 18 (10) (2021) 1122–1127.

[27] 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.

[28] 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, Journal of clinical epidemiology 135 (2021) 79–89.

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

[30] D. Chicco, M. J. Warrens, G. Jurman, The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment, IEEE Access 9 (2021) 78368–78381.

[31] 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 mining 14 (1) (2021) 1–22.

[32] D. Chicco, M. J. Warrens, G. Jurman, The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation, PeerJ Computer Science 7 (2021) e623.

[33] E. Rendón, I. Abundez, A. Arizmendi, E. M. Quiroz, Internal versus external cluster validation indexes, International Journal of computers and communications 5 (1) (2011) 27–34.

[34] 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.

[35] 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.

[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, Proceedings of Machine Learning and Systems 3 (2021).

[37] J. Demšar, Statistical comparisons of classifiers over multiple data sets, The Journal of Machine Learning Research 7 (2006) 1–30.

[38] K. E. Taylor, Summarizing multiple aspects of model performance in a single diagram, Journal of Geophysical Research: Atmospheres 106 (D7) (2001) 7183–7192.

[39] X. Liu, B. Glocker, L. Oakden-Rayner, The medical algorithmic audit, The Lancet Digital Health (2022). doi:10.1016/S2589-7500(22)00003-6.

[40] 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).

[41] 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, European urology 74 (6) (2018) 796–804.

[42] 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 medicine 17 (1) (2019) 1–7.

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

[44] M. Sundararajan, A. Najmi, The many shapley values for model explanation, in: International Conference on Machine Learning, PMLR, 2020, pp. 9269–9278.

[45] H. Chefer, S. Gur, L. Wolf, Transformer interpretability beyond attention visualization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 782–791.

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

[47] A. Holzinger, A. Carrington, H. Müller, Measuring the quality of explanations: the system causability scale (scs), KI-Künstliche Intelligenz (2020) 1–6.

[48] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).

[49] C. Schwartzenberg, T. van Engers, Y. Li, The fidelity of global surrogates in interpretable machine learning, BNAIC/BeneLearn 2020 (2020) 269.

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

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

[52] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, C. Ré, Hidden stratification causes clinically meaningful failures in machine learning for medical imaging, in: Proceedings of the ACM conference on health, inference, and learning, 2020, pp. 151–159.

[53] J. Cowls, A. Tsamados, M. Taddeo, L. Floridi, The ai gambit—leveraging artificial intelligence to combat climate change: Opportunities, challenges, and recommendations, `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3804983` (2021).

[54] B. Van Calster, L. Wynants, D. Timmerman, E. W. Steyerberg, G. S. Collins, Predictive analytics in health care: how can we know it works?, Journal of the American Medical Informatics Association 26 (12) (2019) 1651–1654.

[55] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (1) (2016) 1–9.

[56] N. Barnes, Publish your computer code: it is good enough, Nature 467 (7317) (2010) 753–753.

[57] D. G. Fryback, J. R. Thornbury, The efficacy of diagnostic imaging, Medical decision making 11 (2) (1991) 88–94.

[58] 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, European Radiology (2021) 1–8.

[59] 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, arXiv preprint arXiv:2101.03989 (2021).

[60] 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, in: AMIA Annual Symposium Proceedings, Vol. 2019, American Medical Informatics Association, 2019, p. 1002.