Checklist for supervised clinical ML study

Before paper submission

Study design (Part 1)	**Completed: page number**		**Notes if not completed**
The clinical problem in which the model will be employed is clearly detailed in the paper.	x	5-6	
The research question is clearly stated.	x	5-6	
The characteristics of the cohorts (training and test sets) are detailed in the text.	x	7-8, supplementary materials	
The cohorts (training and test sets) are shown to be representative of real-world clinical settings.	x	7-8, supplementary materials	
The state-of-the-art solution used as a baseline for comparison has been identified and detailed.	x	6, 9-10, 11, 15-16, 18-19	NiChart integrates multiple components for image processing, data harmonization and deriving ML-based imaging biomarkers. We have described state-of-the-art methods for each component in the introduction and described improvements of the methods we used. Detailed comparisons to benchmark methods were presented in corresponding methodology papers cited in the text.
Data and optimization (Parts 2, 3)	**Completed: page number**		**Notes if not completed**
The origin of the data is described and the original format is detailed in the paper.	x	7-8, supplementary materials	
Transformations of the data before it is applied to the proposed model are described.	x	8-10	
The independence between training and test sets has been proven in the paper.	x	9-10, 18-19	All ML models were trained with careful cross validation, with nested cross-validation for parameter optimization. Training data included cross-sectional data, and longitudinal time points were excluded in testing set. Data harmonization experiments were tested using both regular cross-validation and “leave site out” cross-validation
Details on the models that were evaluated and the code developed to select the best model are provided.	x	8-10, 18-19	Details of parameter optimization and model selection for harmonization and machine learning models are presented in detail in

			corresponding method papers cited in the text
Is the input data type structured or unstructured?		x Structured	☐ Unstructured
Model performance (Part 4)	**Completed: page number**		**Notes if not completed**
The primary metric selected to evaluate algorithm performance (eg: AUC, F-score, etc) including the justification for selection, has been clearly stated.	x	9-12	
The primary metric selected to evaluate the clinical utility of the model (eg PPV, NNT, etc) including the justification for selection, has been clearly stated.	x	11-12	
The performance comparison between baseline and proposed model is presented with the appropriate statistical significance.	x	9-12	Methodology papers provide comprehensive validation details for each component. We provided validation results for statistical data harmonization for in-sample and out-of-sample harmonization
Model Examination (Parts 5)	**Completed: page number**		**Notes if not completed**
Examination Technique 1[a]	x	8-10	First part of our examination involved validation of statistical data harmonization of the pooled imaging variables
Examination Technique 2[a]	x	11-12	Second part of our examination involved validation/interpretation of NiChart machine-learning indices using clinical data and disease groups
A discussion of the relevance of the examination results with respect to model/algorithm performance is presented.	x	13-14	
A discussion of the feasibility and significance of model interpretability at the case level if examination methods are uninterpretable is presented.	NA	NA	Final imaging biomarkers derived as part of NiChart are interpretable by design. They are positively associated with aging and disease related brain atrophy
A discussion of the reliability and robustness of the model as the underlying data distribution shifts is included.	x	8-10	Generalizability of statistical harmonization was shown in the methodology paper (Pomponio et al., 2020) using both semi-synthetic and real datasets. We showcased and discussed the replicability of out-of-sample harmonization with

			additional experiments on two datasets
*Common examination approaches based on study type: * For studies involving exclusively structured data coefficients and sensitivity analysis are often appropriate * For studies involving unstructured data in the domains of image analysis or NLP: saliency maps (or equivalents) and sensitivity analysis are often appropriate			
Reproducibility (Part 6): choose appropriate tier of transparency			**Notes**
Tier 1: complete sharing of the code		x	
Tier 2: allow a third party to evaluate the code for accuracy/fairness; share the results of this evaluation		□	
Tier 3: release of a virtual machine (binary) for running the code on new data without sharing its details		□	
Tier 4: no sharing		□	

PPV: Positive Predictive Value

NNT: Numbers Needed to Treat

[a] Common examination approaches based on study type: for studies involving exclusively structured data, coefficients and sensitivity analysis are often appropriate; for studies involving unstructured data in the domains of image analysis or natural language processing, saliency maps (or equivalents) and sensitivity analyses are often appropriate. Select 2 from this list or chose an appropriate technique, document each technique used on the appropriate line above.