

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.	✓	1	
The research question is clearly stated.	✓	1	
The characteristics of the cohorts (training and test sets) are detailed in the text.	✓	5-10	
The cohorts (training and test sets) are shown to be representative of real-world clinical settings.	✓	5-10	
The state-of-the-art solution used as a baseline for comparison has been identified and detailed.	✓	32-33	
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.	✓	23-24	
Transformations of the data before it is applied to the proposed model are described.	✓	32	
The independence between training and test sets has been proven in the paper.	☐		The training and test sets are non overlapping for all setting.
Details on the models that were evaluated and the code developed to select the best model are provided.	✓	31	
Is the input data type structured or unstructured?	☒ 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.	✓	33-34	
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.	☐		We do not evaluate the clinical utility of the model.
The performance comparison between baseline and proposed model is presented with the appropriate statistical significance.	✓	6-10	
Model Examination (Parts 5)	Completed: page number		Notes if not completed
Examination Technique 1 ^a	☐		

Examination Technique 2 ^a	✓	12-14	
A discussion of the relevance of the examination results with respect to model/algorithm performance is presented.	✓	12-14	
A discussion of the feasibility and significance of model interpretability at the case level if examination methods are uninterpretable is presented.	☐		
A discussion of the reliability and robustness of the model as the underlying data distribution shifts is included.	✓		The whole manuscript focuses on the question of robustness and how the model operates under (different amounts and kinds of) distribution shifts.
*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	☐		
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.