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.	☐		The clinical problem is not of interest in our evaluation. We only want to show the algorithm's performance and not interpret clinical parameters.
The research question is clearly stated.	☑		
The characteristics of the cohorts (training and test sets) are detailed in the text.	☑		
The cohorts (training and test sets) are shown to be representative of real-world clinical settings.	☑		
The state-of-the-art solution used as a baseline for comparison has been identified and detailed.	☑		
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.	☑		
Transformations of the data before it is applied to the proposed model are described.	☑		
The independence between training and test sets has been proven in the paper.	☑		
Details on the models that were evaluated and the code developed to select the best model are provided.	☐		Our goal was not to find a best model but rather compare the models that were trained in a federated scenario and a centralized scenario.
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.	☑		
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 did not use any metric to evaluate the clinical utility as this was not the goal of our work.
The performance comparison between baseline and proposed model is presented with the appropriate statistical significance.	☐		This was not the goal of our work.
Model Examination (Parts 5)	**Completed:**	**page number**	**Notes if not completed**

Examination Technique 1[a]	☐		We did not examine the models in detail, as we were only interested in the compared performance between centralized, individual and federated models
Examination Technique 2[a]	☐		
A discussion of the relevance of the examination results with respect to model/algorithm performance is presented.	☐		
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.	☐		
*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		☑	https://github.com/FeatureCloud/evaluation
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.