

Additional File 3: APPRAISE-AI Checklist

No.	Appraisal Item	Score	Explanation [page number of study]
Title			
1	Identify the report as an AI application to a specific clinical question. [Max score 1]	1	
i	The words artificial intelligence, AI, machine learning, deep learning, or other terminology related to artificial intelligence are reported in the title	Y	Title – p. 1.
ii	The outcome of interest predicted by the AI model is reported in the title	Y	Title – p. 1.
iii	The target population in which the AI model will be used is reported in the title	Y	Title – p. 1.
Introduction			
2	Background: Describe the clinical problem and rationale for developing AI models. Review existing relevant literature exploring AI models for the problem being addressed. [Max score 1]	1	
i	The clinical context and rationale for developing/updating an AI model(s) to address the clinical problem are presented	Y	Background – pp. 2–5.
ii	A synthesis of existing AI models that predict the same outcome is provided. If there are no existing models, this should be stated	Y	Background – pp. 3–5; Comparison with Prior Works – pp. 46–49.
3	Objective and Problem: Clearly state what the proposed AI model(s) aims to address with respect to the study population and outcome. [Max score 1]	1	
i	The objectives are presented	Y	Background – pp. 4–5.
ii	The target population and outcome of interest are stated	Y	Background – pp. 4–5.
Methods			
4	Source of Data: Describe how the dataset was obtained (e.g., single/multi-center, local/national database, etc.), and study period. If relevant, the diversity of the dataset is also described (e.g., inclusion of community hospitals, low/middle income populations, and institutions from other countries). [Max score 8]	4	
i	How many institutions were included in the dataset?	Multiple institutions	Methods – Datasets and Experimental Design – pp. 11–14.
ii	Was the study period (start and end dates) reported?	Y	Methods – Datasets and Experimental Design – pp. 11–14.
iii	Was the length of follow-up reported, if applicable?	Not applicable	Not applicable to a retrospective static-frame benchmark without longitudinal follow-up.
iv	What was the setting(s) of the institutions included in the data or inferred based on their description? If not reported or unknown, select No. Academic institutions	Y	Methods – Datasets and Experimental Design – pp. 11–14.
v	Institutions from multiple (> 1) countries	Y	Methods – Datasets and Experimental Design – pp. 11–14.
vi	Community-based or rural hospital(s)	N	Not reported based on source datasets
vii	Low/middle income patient populations	N	Not reported based on source datasets

5	Eligibility criteria: Specify all criteria for inclusion/exclusion of patients and features. Provide appropriate details (e.g., adults, age > 18) and rationale. [Max score 3]	3	
i	Inclusion criteria are provided	Y	Methods – Datasets and Experimental Design – pp. 11–14.
ii	Exclusion criteria are provided	Y	Methods – Datasets and Experimental Design – pp. 11–14.
iii	Details and rationale for criteria are provided	Y	Methods – Datasets and Experimental Design – pp. 11–14.
6	Ground truth: Define the ground truth of interest. Describe how it was collected (e.g., manual annotation by experts) and encoded (e.g., binary, categorical, dichotomized continuous, continuous variable, etc.). [Max score 6]	4	
i	Ground truth of interest is clearly defined. For unsupervised learning, describe what measure(s) and associated data will be used to assess cluster validity (e.g., correlating disease specific features with overall survival)	Y	Methods – Clinical reference framework and scope – pp. 6–9; Datasets and Experimental Design – pp. 11–14; Data integrity layer – pp. 14–17.
ii	How was the ground truth determined?	Single, expert	Partially fulfilled: stage-specific references are heterogeneous. Expert head geometry and landmarks support Find/Measure, but Confirm uses curated boxes and model-assisted provisional masks rather than independent expert full-plane adjudication – pp. 11–17.
7	Data abstraction, cleaning, preparation: Describe the methods used to develop the final dataset, with consideration of feature abstraction, handling of missing data, feature engineering, and removal of features. [Max score 7]	5	
i	Rationale provided for choice of candidate features (e.g., based on prior research, clinical relevance, available data, etc.)	Y	Methods – Clinical reference framework and scope – pp. 6–9; System overview – pp. 9–11.
ii	Time-windows for abstracted features are specified (e.g., vital signs recorded within the past 12 hours will be used to predict sepsis)	Not applicable	Not applicable: ATLAS uses same-frame static imaging inputs rather than longitudinal feature windows.
iii	How was missing data handled? If there is no missing data, it should be clearly stated that there is no missing data, select Not applicable. If it is unclear whether there is missing data or how it was handled, select Not reported	Explicit modeling with justification	Methods – Datasets and Experimental Design – pp. 11–14. Missing or implausible pixel spacing uses a prespecified training-derived fallback with retained provenance.
iv	Transformation/Augmentation: Details provided for how data was altered to change its representation (e.g., normalization, log-transformation, one hot encoding, image rotation, image translation, adjusting image contrast) If not performed, it should be clearly stated that it was not performed, select Not applicable. If it is unclear whether it was performed or not explicitly stated, select No	Y	Methods – Data integrity layer – pp. 14–17; Image preprocessing – p. 17; Confirm stage – pp. 19–21.
v	Modification/Cleaning: Details provided for how data was altered in a non-uniform manner (e.g., outlier removal). If not performed, it should be clearly stated that it was not performed, select Not applicable. If it is unclear whether it was performed or not explicitly stated, select No	Y	Methods – Data integrity layer – pp. 14–17; Measure stage – pp. 21–23.
vi	Outline any methods used to remove features (e.g., clinical judgement, principal component analysis, recursive feature elimination, correlation, or ablation analysis), if applicable	N	Not explicitly reported: no conventional feature-selection or dimensionality-reduction procedure is described, but

	If not performed, it should be clearly stated that it was not performed, select Not applicable. If it is unclear whether it was performed or not explicitly stated, select No.		the manuscript does not expressly state that none was performed.
8	Data splitting: Specify how the data was divided into the training, validation, and testing cohorts. [Max score 7]	7	
i	What was the method of data splitting used?	Held-out validation cohort	Methods – Datasets and Experimental Design, Table 1 – pp. 11–14.
ii	What was the method used to evaluate model generalizability	External validation	Methods – Datasets and Experimental Design – pp. 11–14; External evaluation – pp. 24–26.
iii	Were there any concerns of data leakage (i.e., data preprocessing performed prior to data splitting, training and testing on the same data)?	N	Methods – Datasets and Experimental Design – pp. 11–14; Data Integrity Layer Performance – pp. 32–33. Patient-exclusive partitions, locked testing, subject-exclusive external evaluation, and overlap checks address leakage.
9	Sample size calculation: Provide rationale for sample size required for model development (e.g., based on power calculation). [Max score 5]	0	
i	Minimum sample size required reported	N	Not reported: no minimum required sample size was calculated.
ii	Minimum number of events required reported	N	Not applicable/not reported: conventional outcome-event requirements were not calculated for this continuous biometry benchmark.
iii	Details provided for sample size calculation (e.g. assumptions for event rates, target performance, power, significance level). Can be provided in supplementary material	N	Not reported: no formal power or prospective sample-size calculation was undertaken because fixed public benchmark datasets were used.
10	Baseline: Describe the baseline model that will serve as a comparison for the AI model(s). [Max score 8]	6	
i	Existing model from prior literature used for comparison	Y	Results - Comparison with Prior Works
ii	Regression model using same features in AI model used for comparison	N	Not included: no regression model using the same inputs was evaluated.
iii	Domain expert (e.g., clinician judgement) or current standard of care (gold standard) used for comparison	Y	Methods - Clinical reference framework and scope of computational operationalization - ISUOG standard pp. 5 to 6
11	Model description: Describe the AI model(s) and software libraries investigated. [Max score 2]	2	
i	Type of AI model(s) reported (e.g., random forest, support vector machine, convolutional neural network)	Y	Methods – Find, Confirm, and Measure stages – pp. 18–23.
ii	Software libraries reported (e.g., scikit-learn 1.1.2)	Y	Additional file 1, as referenced under Image preprocessing – p. 17.
12	Hyperparameter tuning: Specify all model hyperparameters that were optimized, the search space for hyperparameter tuning, and evaluation metric(s) used to optimize parameters. Details can be included in Supplementary Material. [Max score 5]	2	

i	Hyperparameters that are tuned are listed (e.g., number of trees, max depth, number of neurons)	Y	Methods – Confirm stage – pp. 19–21; Adaptation and aggregation – pp. 26–29; Additional file 1.
ii	Optimization metric is specified (e.g., accuracy, AUROC, etc.)	Y	Methods – Confirm stage – pp. 19–21; Evaluation metrics – pp. 29–31; Additional file 1.
iii	Hyperparameter search strategy is described (e.g., random-, grid-search, etc.)	N	Not fully reported: development used validation-based and staged/manual choices, but no consistent formal grid, random, or Bayesian search strategy is documented across components.
iv	Search space for hyperparameters are provided	Not reported	Not reported comprehensively: full search spaces are not supplied for all tuned settings, and defaults versus fixed settings are not consistently distinguished.
Results			
13	Cohort characteristics: Provide the total cohort size and summary statistics of the training, validation (if used), and testing cohorts, including incidence of the ground truth of interest. [Max score 4]	4	
i	Total cohort size, number of samples with missing data, and follow-up time (if applicable) are reported	Y	Methods – Datasets and Experimental Design, Table 1 – pp. 11–14; Results – Data Integrity Layer Performance – pp. 32–33.
ii	Summary statistics of each cohort provided to show similarities and differences among cohorts	Y	Methods – Datasets and Experimental Design, Table 1 – pp. 11–14; Discussion – Scope of inference – pp. 57–59.
iii	Incidence of ground truth(s) of interest is reported	Y	Results – Internal Component Performance – pp. 33–36; External Model Transfer – pp. 41–43.
14	Model specification: Present the final AI model(s) and specify the final panel of features included and hyperparameters tuned. Final hyperparameters can be listed in Supplementary Material. [Max score 3]	3	
i	Type of AI model(s) is reported	Y	Methods – Find, Confirm, and Measure stages – pp. 18–23.
ii	Final set of features are reported	Y	Methods – System overview – pp. 9–11; Find, Confirm, and Measure stages – pp. 18–23.
iii	Final set of hyperparameters are reported	Y	Additional file 1, as referenced under Image preprocessing and model-stage methods – pp. 17–23.
15	Model evaluation: List the evaluation metrics used to assess performance and calibration, including the justification for selection. [Max score 5]	3	
i	Measure(s) for model discrimination is reported (e.g., AUROC, AUPRC, c-index, etc.) If multiple measures of discrimination are provided and at least one includes a measure of statistical significance, select Measure(s) with statistical significance	Y	Methods – Evaluation metrics – pp. 29–31; Results – Internal Component Performance and paired comparisons – pp. 33–36.
ii	Rationale provided for which metric is most clinically relevant for the problem at hand	Y	Methods – Evaluation metrics and Joint measurement criterion – pp. 29–31.

iii	Measure(s) for model calibration is reported (e.g., calibration plots, calibration slope and intercept) If both calibration plot and statistical summary of calibration are provided, select Calibration plot	Not applicable	Calibration plots, calibration slope, and intercept are not reported. These are not directly applicable to the continuous biometric errors and evidence states as currently interpreted.
16	Clinical utility assessment: Describe appropriate metrics for readers to understand the risk/benefit trade-offs of using the AI model at the specified decision threshold (e.g., decision curve analysis). [Max score 5]	0	
i	Measure(s) of clinical utility is reported If both sensitivity or specificity for a specified threshold and decision curve analysis are provided, select Decision curve analysis	No assessment of clinical utility	Not within the scope of the study, as the study aims to prepare the technical readiness toward clinical utility study as future work
17	Bias assessment: Compare evaluation metrics for the AI model(s) and reference standard when stratified by patient- and task-specific subgroups to identify subgroups that benefit, are not helped at all, or harmed by the models. Patient-specific subgroups may include age group, gender, ethnicity, or socioeconomic status. Task-specific subgroups are disease-specific and may include risk stratification (e.g., low-, intermediate-, and high-risk disease in prostate cancer), or subtyping (e.g., different bacteria in positive blood cultures). [Max score 6]	2	
i	Patient-specific: Performance (e.g., AUROC) is evaluated across at least one subgroup	N	Not applicable: protected-attribute patient-subgroup performance could not be assessed because demographic metadata were unavailable – pp. 57–61.
ii	Patient-specific: Clinical utility (e.g., sensitivity or specificity for a specified threshold) is evaluated across at least one subgroup	N	Not applicable
iii	Patient-specific: More than one subgroup is evaluated in either performance or clinical utility	N	Not applicable
iv	Task-specific: Performance (e.g., AUROC) is evaluated across at least one subgroup	Y	Results – Confirm evidence as an indicator of measurement reliability – pp. 38–41; External Model Transfer – pp. 41–43.
v	Task-specific: Clinical utility (e.g., sensitivity or specificity for a specified threshold) is evaluated across at least one subgroup	N	Not fulfilled: no task-specific clinical-utility or decision-curve analysis was conducted.
vi	Task-specific: More than one subgroup is evaluated in either model performance or clinical utility	Y	Results – Confirm evidence strata – pp. 38–41; FP versus UCL external-source analyses – pp. 41–45.
18	Error analysis: Analyze predictive errors to identify characteristics that are more prone to inaccurate predictions. Determine if there are any surprise errors (e.g., clearly inaccurate predictions based on clinical judgement). [Max score 4]	4	
i	Analysis of predictive errors is reported	Y	Results – Data Integrity Layer Performance – pp. 32–33; Ablation Analysis – pp. 36–41; External Model Transfer and Adaptation – pp. 41–45.

ii	Analysis of surprise errors is reported	Y	Discussion – Implications – pp. 54–57; Strengths, limitations, and future research – pp. 59–61.
19	Model explanation: Describe methods used to explain AI models (e.g., SHAP, LIME, Grad-CAM) [Not scored]	0	
i	Model explanations are provided	N	Not reported: no SHAP, LIME, Grad-CAM, or comparable post-hoc model-explanation method was used. The structured evidence record is an architectural traceability mechanism rather than a post-hoc explanation method.
Discussion			
20	Critical analysis: Describe main findings and limitations of the study. [Max score 5]	5	
i	An overall interpretation of the results is presented, which may include: - New predictors of the ground truth of interest discovered using AI - Strengths of the AI model(s) compared to current models in the literature - Why the AI model(s) performed better/worse than what is currently available? - (Optional) If feature importance rankings were used, describe whether they were aligned with clinical intuition and known prognostic factors	Y	Discussion – Interpretation of the key findings – pp. 49–54; Comparison with Prior Works – pp. 46–49.
21	Implementation into clinical practice: Describe how the AI model(s) can be applied to clinical practice, with respect to the potential to improve patient care, clinical decision-making, and/or efficiency. [Max score 1]	1	
i	Potential application(s) to clinical practice and future directions are discussed	Y	Discussion – Implications – pp. 54–57; Strengths, limitations, and future research – pp. 59–61.
22	Limitations: Discuss the limitations of the AI model(s), with consideration of the data, features, model(s), and/or biases. [Max score 2]	2	
i	Limitations are discussed	Y	Discussion – Scope of inference – pp. 57–59; Strengths, limitations, and future research – pp. 59–61.
Other information			
23	Disclosures: Disclose all financial relationships, sources of funding, and potential conflicts of interest. [Max score 1]	1	
i	All relevant disclosures are reported	Y	Declarations – Competing interests and Funding – p. 64.
24	Transparency: Share the data, data dictionary, source code, or release an application that runs the code. [Max score 10]	7	
i	Data dictionary: A description is provided for all features and ground truth, with consideration of the following: - Data type (i.e., categorical or numerical) - Method of collection or measurement (e.g., serum hemoglobin in g/dL) - Range of values (e.g., yes or no, 0.5-250 g/dL)	Y	Methods – System output – pp. 23–24; List of abbreviations – pp. 63–64. The principal input, reference, output, units, and provenance fields are described.

ii	Data availability: How can other researchers access the data used in the study? Data availability needs to be explicitly stated to receive points	Established repository	Declarations – Availability of data and materials – p. 64.
iii	Model availability: How can other researchers access the model(s) used in the study?	Y	Declarations – Availability of data and materials – p. 64.
iv	Nomogram/scoring system/website available to use model for single predictions	Y	Declarations – Availability of data and materials – p. 64.
v	Trained model available to generate prediction in bulk (i.e., from a dataset)	Y	Declarations – Availability of data and materials – p. 64.
vi	Complete source code available	Y	Declarations – Availability of data and materials – p. 64.
vii	Executable end-to-end (e.g., dependency file, documentation on how to run the code) available	Y	Declarations – Availability of data and materials – p. 64.

Overall APPRAISE-AI score (out of 100)	68
Quality based on overall APPRAISE-AI score	High quality
Clinical Relevance (out of 4)	4
Data Quality (out of 24)	16
Methodological Conduct (out of 20)	13
Robustness of Results (out of 20)	9
Reporting Quality (out of 12)	12
Reproducibility (out of 20)	14