Table S1. HUMANE-Collaborators from Five Countries

Australia
Anuroop Gaddam
Jacqueline Boyle
Sandeep Reddy

India
Thanga Prabhu

Sweden
Toomas Timpka

United Kingdom
Stefanie Lip
Christopher Sainsbury
Gabriel Reines March
Chris Carlin
David J Lowe
Linsay McCallum
Shane Burns
Clea du Toit

United States of America
Shashi Tripathi
Salim Surani
Mack Sheraton
Ashish K Khanna
Ravi Kiran Bhaskar
Nitesh Jain
Chris Aakre
Kamal Maheshwari
Jacek Cywinski
Vitaly Herasevich
Mohammad Bilal
Shekhar Ghamande
Sravanthi Parasa
Vishwanath Pattan
Allon Kahn
Neha Deo
Khalid Moidu
Shyam Visweswaran
Piyush Mathur
Chaitanya Mamillapalli

Table S2. Checklist questions and their individual score

Checklist Questions	MEAN (+/- SD)	MEDIAN (25-75% IQR)
Section 1: Title		
Is the title relevant to research in the field of Artificial Intelligence/Machine Learning in Medicine? Yes/No	4.5 (+/-0.71)	5 (4-5)
Does the title align with any of the following terms or "related terms": Artificial Intelligence, Machine Learning, or Deep Learning? Yes/No	4.2 (+/-0.90)	4 (4-5)
Section 2: Abstract		
Does the abstract provide a summary of the following: objectives, study design, setting, target population, statistical analysis, results, and conclusion pertinent to AI/ML in healthcare? Yes/No	4.5 (+/-0.67)	5 (4-5)
Section 3: Introduction		
Has the study provided a background in the context of a clinical domain and the role of Artificial Intelligence in the field? --> Yes/No	4.9 (+/-0.62)	5 (4-5)
Has the author described the background/introduction section as a rationale for the need for research using the following parameters: Why this topic is important: ex. cost/life/time/process savings? -> Yes/No, What is already known in this field? -> Yes/No, What is the knowledge gap? -> Yes/No, What did the authors want to do/did to fix this knowledge gap? -> Yes/No	4.5 (+/-0.62)	5 (4-5)
Has the study defined the objectives and highlighted the scope in the validation or development of the AI model? --> Yes/No	4.5 (+/-0.87)	5 (4-5)
Does the background provide information on the following? Description of current knowledge gap on this topic --> Yes/No - Description of gap in triage or diagnostic pathway --> Yes/No	4.3 (+/-0.89)	5 (4-5)
Which of the following domain(s) has this study explored the potential impact of this model?--> 1. Early Diagnosis 2. Improved Diagnosis 3. Allowed personalized/targeted treatment 4. Prevent/reduce hospital admissions 5. Improve survival 6. Other (check one)	4.0 (+/-1.2)	4 (3-5)
Section 4a: Methods (Data Source)		
Was the study methodology and study pre-specified in terms of the study design (eg: Retrospective/Prospective, Derivation/Validation, Supervised/Unsupervised/Deep ML), including characteristics of the data type collected? --> Yes/No	4.6 (+/-0.55)	5 (4-5)

Is the study timeline specified in terms of the initiation of data collection/model development and the end date of the completed (or ongoing) data collection/model validation? --> Yes/No	4.3 (+/-0.81)	5 (4-5)
Is the dataset obtained from within the intended stage in the care pathway? --> Yes/No	4.1 (+/-0.82)	4 (4-5)
Were the key data pre-processing/pre-curation steps explained? --> Yes/No/Not applicable	4.2 (+/-1.0)	5 (4-5)
Does the disease probability in the dataset differ from the setting in which the model will be deployed? --> Yes/No/Data not available	4.3 (+/-0.92)	5 (4-5)
Is there sufficient clarity on how the data were split/categorized in terms of the Training set, Tuning set, Internal Validation set, and External validation set: --? Very unclear to Highly Clear (1-5 scale)	4.5 (+/-0.87)	5 (4-5)
Are all 3 cohorts clearly defined for model development? (training, validation, and test cohort) --> Yes/No/Not applicable	4.5 (+/-1.1)	5 (5-5)
Were ethical considerations made in ensuring reliable data collection, along with de-identification of patient records, if applicable? --> Yes/No/Not applicable	4.2 (+/-1.1)	5 (4-5)
Section 4b: Methods (Participants)		
Has documented consent been taken from the participants involved in the prospective/intervention study? --> Yes/No/Not applicable	4.3 (+/-0.81)	5 (4-5)
Is there a pre-defined inclusion and exclusion criteria for different model/study cohorts? --> Yes/No/Not applicable	4.5 (+/-0.79)	5 (4-5)
Section 4c: Methods (Outcomes)		
Was the outcome proposed by the AI model well-connected with written methods in the following questions? --> 1. Were the outcomes generated by the AI model in relation to the data sample being assessed? 2. Was the outcome compared with the same standard reference as the training set for sensitivity and specificity? 3. Has the study described a multivariable prediction model, including its role in the assessment of outcomes? --> All 3 Yes/No	4.4 (+/-0.71)	5 (4-5)
For models focused on addressing the knowledge gap (discovery studies), indicate the questions that have been addressed from the following: 1. What is the knowledge gap in this field? 2. What is the reason for the existence of this	4.2 (+/-0.88)	4 (4-5)

knowledge gap? 3. What aspect of this knowledge gap is the model trying to address? --> Yes/No		
For models focused on addressing the triage or diagnostic pathway, indicate the questions that have been addressed from the following: 1. What is the intended role of this model (triage or diagnosis)? 2. Will the model be used as an isolated test or in combination with other diagnostic elements? --> Both Yes/No	4.1 (+/-1.1)	4 (3-5)
Is the experimental protocol designed to prevent overfitting? --> Yes/No	4.1 (+/-1.2)	5 (4-5)
How have they developed the experimental protocol to prevent overfitting? --> 1. Independent training and test validation 2. Cross-fold validation 3. Leave one out of validation 4. Other 5. Not Applicable --> Select one	4.0 (+/-1.3)	4 (4-5)
Section 4d: Methods (Statistical Analysis)		
Has the study pre-specified a statistical analysis plan? --> Yes/No	4.1 (+/-1.1)	4 (4-5)
Has the study specified a range of statistical measures used to compare the Accuracy/ Precision/ Sensitivity/ Specificity of the proposed model? --> Yes/No	4.4 (+/-1.0)	5 (4-5)
Has the study described the predictor model using a internal validation technique? --> Yes/No/Not applicable	4.3 (+/-0.98)	5 (4-5)
Section 5a: Ground Truth (Labels)		
Were the ground truth labels manually determined by experts? --> Yes/No	4.4 (+/-0.99)	5 (4-5)
Were the ground truth labels automatically generated? --> Yes/No	4.2 (+/-1.0)	5 (4-5)
Were any ground truth labels missing? --> Yes/No	4.2 (+/-1.1)	5 (4-5)
Were the ground truth labels added? Prospectively or retrospectively	4.3 (+/-0.78)	4 (4-5)
On a scale of 1-10, how accurate are the ground truth labels? 1-based on a single element, 10-based on a hard outcome (eg: death)	4.1 (+/-1.0)	4 (4-5)
Section 5b: Ground Truth (Expert(s) Review) Section 5b: Ground Truth (Expert(s) Review)		
Which of the following is applicable for the number of experts involved in the review: 1. Single 2. Multiple Independent 3. Use of Adjudicator(s)	3.9 (+/-1.1)	4 (4-5)
Which of the following is applicable regarding the qualification of the expert(s) in the review: --> 1. Sub-specialist with	3.8 (+/-1.0)	4 (3-5)

experience 2. Board-certified specialist 3. Specialist in the domain without sub-specialty accreditation (choose any)		
Was there sufficient availability of clinical information to the expert to make the diagnosis? --> Yes/No/Data not available	4.2 (+/-0.82)	4 (4-5)
Was an inter-observer agreement presented? --> Yes/No/Not applicable	4.2 (+/-0.87)	4 (4-5)
Was there a pre-specified threshold for inclusion of cases where there is non-consensus? --> Yes/No	4.1 (+/-1.1)	4 (4-5)
Section 6: Results		
Has the study described key demographics/characteristics of the cohorts? (age, gender, chronic co-morbidities, patient type, etc.) --> Yes/No	4.3 (+/-0.96)	5 (4-5)
Was model validation performed robustly using out-of-sample external validation test dataset? --> Yes/No	4.4 (+/-0.99)	5 (4-5)
Has the study identified any differences between the development and validation data sets in inclusion criteria, model outcome, and predictors? --> Yes/No	4.3 (+/-1.0)	5 (4-5)
Was the validation dataset distinct? 1. temporally, 2. geographically, or 3. both 4. None?	4.2 (+/-0.85)	4 (4-5)
Has the study mentioned the inclusion and exclusion of data, including the missing data with appropriate justification and/or flow diagram? --> Yes/No	4.5 (+/-0.94)	5 (4-5)
Has the study reported any discrimination measures of performance? --> 1. Accuracy 2. Sensitivity / Recall 3. Specificity 4. Precision 5. ROC curve 6. Precision recall (PR) curve 7. Other (check one)	4.6 (+/-0.75)	5 (4-5)
Has the study reported any calibration measures of performance? --> 1. Calibration plot 2. Hosmer-Lemeshaw test 3. Excepted calibration error 4. Brier score 5. Mean square error (MSE) 6. Other (Check one)	3.9 (+/-1.0)	4 (3-5)
Has the study evaluated model fairness? (Ex. performance is reported for separate sexes) --> Yes/No	3.9 (+/-1.1)	4 (3-5)
Do the training and validation datasets represent the complete spectrum of diagnostic cues for the target population? --> 1. Disease prevalence in the internal validation test dataset representative of the target population in the real world 2. Presence of under or overrepresented subgroups within the training dataset 3. Authors have applied any inclusion or exclusion criteria that create a selection bias 4. Authors have	3.9 (+/-1.3)	4 (3-5)

applied a sampling method (i.e., random sampling) to reduce the risk of spectrum bias? 5. Other (Check one)		
Has the study applied any of the following methods to address class imbalance? 1. Oversampling – adding copies of underrepresented class 2. Under-sampling – removing copies of overrepresented class 3. Replicate the class distribution in the validation test set 4. Other (Check one)	3.9 (+/-1.2)	4 (3-5)
Does the study provide differential diagnoses and confidence estimates? --> Yes/No/Not applicable	4.0 (+/-0.98)	4 (4-5)
Has the study reported the values of the measured variable? --> Yes/No	4.2 (+/-0.93)	4 (4-5)
Has the study compared their results with existing literature, by supporting or challenging their findings? --> Yes/No	4.3 (+/-1.0)	5 (4-5)
Has the study reiterated the purpose of AI technology and what it measures? --> Yes/No	4.2 (+/-1.1)	5 (4-5)
Section 7: Discussion		
Has the study provided a concise summary of their primary result findings on this topic, covering at least 1-3 main points with no/minimal numerical values? --> Yes/No	4.6 (+/-0.78)	5 (4-5)
Has the study evaluated their results with studies in favor/against from previous literature? --> Yes/No	4.3 (+/-0.92)	5 (4-5)
Has the study listed at least 3-5 strengths of their research? --? Yes/No	4.1 (+/-1.1)	4 (4-5)
Has the study listed at least 1-3 weaknesses of their research? --> Yes/No	4.4 (+/-0.71)	5 (4-5)
Has the study paraphrased the first paragraph of their conclusion, tying it with the study title? --> Yes/No	3.7 (+/-1.3)	4 (3-5)
7: Discussion (Other)		
Has the study listed their conflict of interest(s)? --> Yes/No/Not applicable	4.8 (+/-0.48)	5 (5-5)

Table S3. HUMANE Checklist

Section/Topic		Checklist Item	Yes / No / Partial / NA
Section 1: Title			
	1	Is the title relevant to research in the field of Artificial Intelligence/Machine Learning in Medicine?	
	2	Does the title align with any of the following terms or related terms: Artificial Intelligence, Machine Learning, or Deep Learning?	
Section 2: Abstract			
	3	Does the abstract provide a summary of the following: objectives, study design, setting, target population, statistical analysis, results, and conclusion pertinent to AI/ML in healthcare?	
Section 3: Introduction/Background			
	4	Does the background mention the importance of this topic: ex. cost/life/time/process savings?	
	5	Does the background mention what is already known in this field?	
	6	Does the background mention the knowledge gap in this field?	
	7	Does the background mention how the authors aim to fix this knowledge gap?	
	8	Has the study defined the objectives including validation or development of AI/ML?	
	9	Has the study explored any of the following domain(s) for the potential impact of this model? (Triage, Early Diagnosis, Improved Diagnosis, Personalized treatment, Prevent/reduce hospital admissions, Improve survival)	
Section 4a: Methods (Data Source)			
	10	Are the study methodology and design pre-specified (ex: Retrospective/Prospective, Derivation/Validation, Supervised/ Unsupervised/ Deep ML), including characteristics of the data type collected?	
	11	Does the study timeline specify the initiation of data collection/model development and the end date of the completed (or ongoing) data collection/model validation?	
	12	Is the dataset obtained from within the intended stage in the care pathway?	
	13	Are the key data pre-processing/pre-curation steps explained?	
	14	Is the dataset appropriate for the healthcare conditions studied?	
	15	Is there sufficient clarity on dataset for model development (Training/Test/Validation)?	
	16	Is the validation dataset distinct from testing and training datasets?	
	17	Is it explicitly mentioned that the study is compliant/exempt with local ethical committee/IRB/patient privacy/data security regulations?	
Section 4b: Methods (Participants)			
	18	Is there consent taken from the participants involved in the prospective/intervention study?	
	19	Is there a pre-defined inclusion and exclusion criteria for different models/datasets?	
ion 4c: Methods (Outcomes)			
	20	Is the outcome tested by the AI model aligned with written methods?	
	21	Is the distribution of outcomes similar in all training, testing, and validation datasets?	
	22	Is there a description of other multivariable prediction models?	
Models focused on Triage or Diagnostic pathway			
	23	Does the study state the intended role of this model (ex: triage or diagnosis)?	
	24	Does the study state if the model was used as an isolated test or in combination with other diagnostic elements?	
Section 4d: Methods (Statistical Analysis)			
	25	Is there a pre-specified statistical analysis plan?	
	26	Is there a specified range of statistical measures used to compare the Accuracy/ Precision/ Sensitivity/ Specificity of the proposed model?	
	27	Is there a description of the predictor model using an internal validation technique?	
	28	Is the experimental protocol designed to prevent overfitting?	
Section 5a: Ground Truth (Labels)			
	29	Does the study state if ground truth is applicable to the supervised learning method?	
	30	Are evidence-based details provided on the ground truth labeling process?	

	31	Are the ground truth labels manually determined by experts?	
	32	Are the ground truth labels automatically generated?	
	33	Are any ground truth labels missing?	
	34	Does the study state how the ground truth labels were added (prospectively/retrospectively)?	
Section 5b: Ground Truth (Expert(s) Review)			
	35	Is there any mention of a pre-specified threshold for the inclusion of cases where there is non-consensus?	
Section 6: Results			
	36	Has the manuscript described the key demographics/characteristics of the cohorts? (age, gender, chronic co-morbidities, patient type, etc.)	
	37	Is model validation presented using an out-of-sample external validation dataset?	
	38	Has the manuscript presented any difference between the training, testing, and validation data sets in inclusion criteria, model outcomes, and predictors?	
	39	Has the manuscript described either in text or by a flow diagram the impact of applying stated inclusion/exclusion criteria on the final sample size?	
	40	Has the manuscript reported any discrimination measures of performance? (Accuracy, Sensitivity/Recall, Specificity, Precision, ROC curve, Precision Recall (PR) Curve)	
	41	Has the manuscript reported any calibration measures of performance? (Calibration plot, Hosmer-Lemeshaw test, Expected calibration error, Brier score, Mean square error)	
	42	Has the study evaluated algorithmic bias? (Example: for gender, race, ethnicity, socioeconomic status, etc.)	
	43	Are there steps reported to support external validity of other results?	
	44	Has the study applied any of the following methods to address class imbalance? (Oversampling, Undersampling, Replication in validation dataset)	
Section 7: Discussion			
	45	Has the manuscript provided a succinct summary of their primary result findings?	
	46	Has the manuscript compared their results with existing literature, by supporting or challenging their findings?	
	47	Has the manuscript mentioned the strengths of their research?	
	48	Has the manuscript mentioned weaknesses of their research?	
	49	Have the authors provided a justifiable conclusion based on the results presented with a take-home message and implications of the results?	
Section 7: Discussion (Other)			
	50	Have the authors listed their conflict of interest(s)?	