

DECIDE-AI

Proposed new items (participants)

15.04.21

How to read this document

Two members of the research team (Myura Nagendran and Baptiste Vasey) reviewed in parallel the new items proposed by the participants. For each proposed item, one of five actions was taken and is reported in the table here below:

- Do nothing – the main idea of the proposition is already covered in an existing item
- Add the main idea of the proposition to the comments on an existing item of the original list and consider it as part of the Round 1 comment analysis.
- Add the main idea of the proposition to the provisory explanation of an item (not displayed to the participants at this stage)
- Add the main idea of the proposition to an existing item by modifying the item's wording
- Select the proposed item as a new item

After each mention of an item, clarification whether the number refers to the original (Round 1) or updated (Round 2) item list is provided.

Proposed new items

Proposed item (participants)	Justification (participants)	Action taken (research team)
Title (n=7)		
With the goal being xyz to solve the xyz problem	this is a best practice and provides clarity from the start about intent and problem (relevance)	Newly included in item 1a (updated list)
Should say what design of study is - prospective cohort etc	compatible with other recommendations	Added to item 1 comments (original list)
Identify core study methodology and design e.g. "qualitative observational study", "mixed methods", "interviews with clinician users".	Needs to be clear when screening titles for inclusion in systematic reviews.	Added to item 1 comments (original list)

Specify which decision is being impacted and whether it is a standalone performance assessment of user performance		Added to item 1 comments (original list)
Identify the location (internal, external) and timeframe (retrospective, prospective) of an AI system model evaluation study.		Added to item 1 comments (original list)
Identify the process where the algorithm is used	i.e. what clinical field is addressed (for example screening mammography or selection of patients for hemihepatectomy)	Newly included in item 1a (updated list)
Identify the goal of the use of the algorithm	I.e. faster evaluation, higher sensitivity, lower mortality due to better selection	Newly included in item 1a (updated list)
Abstract (n=4)		
Recommend using the SPIRIT/CONSORT-AI items here - but should include type of AI	As above	Added to item 2 comments (original list)
Provide a summary of discrepancy between reported in-silico performance, particularly with respect to patient subpopulations.	If a model does not demonstrate benefit and/or efficacy on certain subpopulations but provides sufficient benefit to qualify for a large-scale trial, such discrepancy and potential causes should be reported.	Added to item 2 comments (original list)
Clear safety and performance metrics	This is important in order to avoid generic claims	Added to item 2 comments (original list)
Currently very broad: - purpose -setting -ethical approval -status of algorithm used (in-house / commercial) -data fed into algorithm -primary outcome measure tested -results on primary outcome measure - conclusion		Added to item 2 comments (original list)
Introduction (n=10)		
Describe the algorithm itself (including type of AI etc)		Covered in item 9 (updated list)

Annex V-8

What is the ultimate expected use of the CDSS; will it replace humans in routine use and only require "validation" which often is not very relevant	cfr section 1 human in the loop is often referred to but not real if the algorithm is very well performing and substitutes humans in majority of routine decisions	Covered in item 3 (updated list)
Golden standard (if available)	Describe the current state-of-the-knowledge on the golden standard the algorithm wants to classify/predict explicitly	Added to item 23 comments (original list)
Describe the action that should be taken (intended use) based on the outcome of the algorithm and its associated impact in terms of patient care.	Explicitly mentioning the action that is associated with the algorithm outcome is important for evaluating its clinical effectiveness and correct use.	Added to item 4 comments (original list), covered by item 10c (updated list), added to item 10c explanation (updated list)
Clearly state the clinical problem that the implementation of the AI system is intended to address	I would be interested for the introduction to clearly state the clinical problem that AI is attempting to solve - i.e. is it excess human variability, workforce shortage, suboptimal human diagnostic performance, lack of widely available expertise, long waiting lists, need to improve cost-effectiveness, desire to improve efficiency etc? This sets up the whole rest of the study with the clinical context needed to evaluate its potential.	Added to item 3 comments (original list), newly included to item 2 (updated list)
Declaration whether the algorithm is a medical device.	If the algorithm is a medical device then the associated regulations may impose additional safety and performance requirements that may not necessarily be considered.	Added to item 7 comments (original list), newly included to item 4 (updated list)
State of the art on the same topic	it's important to give the reader some information to know if the proposed method is an advancement/innovation etc	Covered in item 2 (updated list)

To what extent the AI-based decision system is introspectable or understandable by humans or GDPR-compliant (EU law)	the reason for this suggestions is that explainability is attracting increasing attention by stakeholders, policy-makers and politicians	Covered in item 23e (updated list), mention about transparency in the EU Ethics guidelines for trustworthy AI added to item 23e (updated list)
Item 7a. State the study hypothesis in terms of the expected clinical use and benefits of the algorithm.	Elaborating an hypothesis could help explicit how the study could help fill the gap described in point 7 and justify the aim and methods.	Covered in item 3 (updated list)
State that there is considering to Human Factors given.	This will help the user to understand that HF has been considered in the study.	Added to item 2 comments (original list)
Methods (n=21)		
Clarify whether any iterative modification of the algorithm during the study was conducted. If so, provide details of the methods used (eg using a formal Quality Improvement approach)	It is generally recognised that the human-machine relationship of AI needs to be understood before a clinical trial against a non-AI assisted human-clinician led process can be performed in a reliable and valid way. It follows from this that discovery of connections between how the system presents information to the user and how the user reacts in terms of trust is highly likely. The obvious need, when this occurs, is to identify and trial a modification of the interface. An efficient approach to doing this is likely to involve iterative modification and rapid evaluation of outcome change. If this strategy is adopted it is clearly absolutely necessary to explain what was changed, why, and when.	Covered in item 14 and 19 (updated list)

Annex V-8

Evaluate how user perceives understandability and relevance of cdss suggestion for problem at hand	critical factor in accepting or overruling suggestion	Covered in item 23e (updated list)
Describe how the recommendations of the algorithm were presented, including whether explanations were given.	I miss a certain focus on potential explanations in order to make the algorithmic output more explainable. They can be as important (if not more important) as the prediction itself.	Added to item 15 comments (original list), newly included in item 10f (updated list)
Describe any instructions given to participants.	Note: this is different to training. It is about instructions specific to the study.	Added to item 10c explanation (updated list)
Registration number and name of registry	not all studies will have a protocol, but they should be registered...	Added to item 9 comments (original list), newly included in item 6a (updated list)
State whether the study was review by a research ethics/institutional review committee.	This is a standard reporting item for clinical trials. It is related to item 18 as well as PHI safeguards are often the purview of health custodians overseen by these boards.	Added to item 9 comments (original list), newly included in item 6a (updated list)
Describe any digital, physical or informational materials used in the delivery of the CDS algorithm, including those provided to participants or in training of users. Provide information on where the materials can be accessed (such as online appendix, URL). Describe each of the procedures, activities, and/or processes used.	Reproducibility / understanding context; the way a DCSS is implemented can influence outcomes.	Added to items 15 and 24 comments (original list), added to item 8c explanation (updated list)

There should be an item about the safety/data safety or security of the algorithm and it's interface software/hardware. How was this maintained or assessed?	The issue of how the software and algorithm updates, and how this might influence safety or performance. If it is happening dynamically, how is this influencing safety, validity, reliability, etc? Also, how this may influence how the algorithm can be evaluated in a clinical trial What about issues of security, data security etc?	Safety covered in item 13a, 13b, 21a and 21b (updated list), data safety covered in item 6b (updated list)
Describe when and where the study was registered and describe and justify and deviations from the pre-registration.	To allow proper adherence to the guidelines and to facilitate their adoption the items should include as a prerequisite the pre-registration of the study, and here the adherence to the pre-registration as well. In the end, it is highly unlikely that studies will have included all relevant reporting guideline items unless the reporting guidelines are used already during study design as a template.	Added to item 9 comments (original list), point about registration newly included to item 6a (updated list), point about protocol deviation added to item 6a explanation (updated list)
Do you have actively reduced variation in subgroup performance?	This should be built-in from the start.	Added to item 51 comments (original list), point about subgroup performance partially covered in item 22 (updated list)
Item 14a. Describe the deployment of the algorithm within the existing clinical infrastructure.	Knowing the actions required to install the algorithm informs on the reproducibility of the study.	Newly included to item 10d (updated list)
Explain any known bias or lack of diversity within the data including whether data was sourced from real world, clinical trial or synthetic	Algorithm may less accurate for sub-groups of differing sex, ethnicity, age or with certain co-morbidities	Covered by item 8a and 10e

The impact of any iterative improvement of the algorithm or interface on the performance of the algorithm should be presented using the agreed outcome measures, and specifying any additional measures used to complement these.	Clearly it is important, if iterative modification occurs, to show as clearly as possible what difference it made to outcomes.	Covered in item 19 (updated list)
Methods reporting should specify any a priori identification of patient and user subgroups which are of interest because either empirical findings, background knowledge or theory suggests that they are likely to have values for user trust of the machine which are significantly different from other groups, or which evolve differently over time. Pre-specification for subgroup analysis avoids any suspicion of data-dredging while providing important information for design of an RCT both in terms of inclusion criteria and subgroup analysis.		Added to item 12 explanation (updated list)
Describe any efforts to address potential sources of bias	This may not apply to all study types, but where the study is observational, STROBE recommends considering bias in the Methods	Added to item 47 comments (original list), partially covered in item 25 (updated list)
For each user (such as radiologist, MD, nursing assistant), describe their expertise, background, and any specific training given.	Expertise of user may influence results.	covered in item 8c and 17b (updated list)

Describe the type(s) of location(s) where the algorithm was used, including any necessary infrastructure or relevant features.		Covered in item 10a (updated list)
If the intervention was modified during the course of the study, describe the changes (what, why, when, and how)		Covered in item 19 (updated list)
If algorithm adherence or fidelity was assessed, describe how and by whom, and if any strategies were used to maintain or improve fidelity during the course of the study, describe them.		added to item 31 comments (original list), point about strategies to improve fidelity added to item 10c explanation (updated list), point about adherence assessment partially covered in item 14 (updated list)
Unit of assignment (the unit being assigned to use the algorithm, e.g., individual, centre, community); unit of analysis if different from unit of assignment		added to item 12 comments (original list), added to item 7 explanation (updated list)
Not certain how to word this, but was the algorithm or it's interface or hardware undergoing updates (self initiated) during the trial?	Was this or can this be checked? This could influence the performance of the algorithm and its output, and certainly data safety and security. Also, would create a more complex situation, as the algorithm would have been evolving in a way while being used and it's effect of human behaviour evaluated.	Point about updates newly included to item 19 (updated list)

Results (n=7)		
Describe the overall performance of the algorithm in numeric terms, in the context of the clinical pathway in question.	Although Qs 32, 33, 34 address this, these 3 questions are not always applicable, and the user may be left without a key performance measure. The performance measure is important for ML applications as it supports an appraisal of when and where to use it clinically, and how acceptable the algorithm is. Performance can change in a real world context and such an in silico measure is insufficient.	Added to item 32 comments (original list)
Report on algorithmic bias by socioeconomic and baseline demographic features.	It is important to understand if disproportionate bias is present and if any subgroups of patients could be harmed by the AI algorithm.	Generic point about subgroup selected as new item 22 (updated list), specific point about socioeconomic and baseline demographic added to item 22 explanation (updated list)
Report on effect of the AI system by user group (i.e. years' training, job title etc).	The effect of the AI system is likely to be quite different depending on the clinical expertise and personality of the user. This must be teased out in the reporting of the results.	Generic point about subgroup selected as new item 22 (updated list), specific point about user groups added to item 22 explanation (updated list)
Describe any instances where users decided to change their mind based on the algorithm's recommendation.	The instances where users actually change their decision making are essential to assess the added value of the algorithm. When there are no instances, the algorithm may be superfluous. When all these instances lead to erroneous conclusions, the algorithm may actually be harmful, etc.	Newly included in item 23a

Annex V-8

Describe the equipment used to acquire the data that was analysed by the AI system	The precise equipment used to acquire images/laboratory values/EHR data/physiological data/genetic data (etc) is crucial for others to consider future implementation in their setting. I don't think this has been added so far.	Added to item 16 comments (original list), added to item 10e explanation (updated list)
Describe any analyses evaluating the independent effect of explainability methods and their influence on decision-making.	Different types of explainability may have different, independent influences on clinician decision-making. Their influence on the overall decision-making picture should be evaluated in parallel to the model.	added to item 22 explanation (updated list)
Report the ways in which patient feedback was obtained and incorporated into the study's design and conduct.	Even if patients do not understand the algorithm itself, how patient feedback was solicited and incorporated is still very important. Patients' feedback may have been purely around the intervention but that would still be relevant here.	Newly included in item 15 (updated list)
Discussion (n=9)		
Describe the benchmark aim for the overall performance of the algorithm in numeric terms, and how that was arrived at	Experts will often repeat that they want AI algorithms to perform at least as well as human experts. However, clinically, this is not always the right litmus test, and in any case it requires quantification and thought in each clinical context.	Added to item 24 explanation (updated list)
Demonstrate conformity to ISO 62366	This is a mandatory regulatory process for Class II and above devices, and therefore should be reported.	Added to item 26 explanation (updated list)

Annex V-8

Analyse and report performance metrics of the model over time.	When data-collection protocols or other external changes outside of the model ecosystems happen, ML model performance can deteriorate. Whether such events were identified or tested for during the small-scale evaluation is important to know and look for before a large-scale trial is conducted.	Newly included in item 20a (updated list)
Describe model maintenance, retraining, auditing protocol in detail.	Describe when and how such a model will be updated, the frequency, and what data it will take in. Ideally a model that undergoes any update should go through the entire life-cycle of i) in-silico evaluation, ii) small-scale trial, iii) large scale trial. If models will be updated using data generated after treating the patients using the model itself, the underlying shift may have unintended effects on model behaviour. Ideally such updates should be discouraged.	Newly included in item 19 (updated list), added to item 19 explanation (updated list)
Discuss strengths and limitations of study, and implications of limitations for interpretation of findings.	This is a standard item for the Discussion of any empirical study but it is nevertheless usually included in reported standards.	Selected as new item 28 (updated list)
Compare findings with findings from similar studies.	This is a standard item for the Discussion of any empirical study but it is nevertheless usually included in reported standards.	Added to item 24 and 28 explanation (updated list)

Annex V-8

Discuss how the population used for this study compares to the population used to perform the initial training and validation of the algorithm.	It's very interesting to know how the population differs (i.e. is it the same geographical site, same equipment, same clinical workflow etc), or is it a true external evaluation? Useful to re-state that there is no overlap in unique patients included in this study vs any previous research. This information will affect how strongly the results can be interpreted.	Added to item 6 comments (original list), partially covered by item 9 (updated list)
Describe need / possibility for re-calibration / model re-training in the future	Model performance likely to change over time (either by Hawthorn effect or because of other changing practice) so needs to be a discussion of how this might be evaluated / likely impact.	Covered in item 27 (updated list), added to item 27 explanation (updated list)
Discuss how errors of the system relate to the frequency and severity of human errors	It should be about the benefit risk ratio, not about errors perse.	Added to item 47 comments (original list), human errors covered in item 21e (updated list), benchmark aspect covered in item 24 (updated list)
Statement (n=6)		
Disclose individual authors contributions to the study.	Contributions of the authors to any study varies depending upon the study and author backgrounds. Funding and conflict of interests are many a times related to the same. The statement provided should clarify who participated in which aspect of the study.	Already standard requirement in most journals
Any involvement of commercial companies in the study design or study itself.	Transparency of the involvement of commercial companies. This can cause biases in the study.	Added to item 53 comments (original list), newly included in item 29 (updated list)
Acknowledgements (if applicable)		Already standard practice in most journals

Annex V-8

Describe any ethics methodology, consultation, or involvement in the design and conduct of the study (e.g., algorithmic fairness, social impact assessment, community consultation).	How researchers address ethical issues is an under-recognized but important aspect of trial conduct. Citation in support: Anderson, J., Eijkholt, M. & Illes, J. Ethical reproducibility: towards transparent reporting in biomedical research. Nat Methods 10, 843-845 (2013)	Selected as new item 16 (updated list)
Re-word 54. Disclose code and data availability as "Disclose code and data availability where possible and whether collaboration possible"	It may not be feasible to disclose code / data for commercial reasons. This is clearly not idea but likely to be reality. Should encourage other avenues for collaborative evaluation of such technologies.	Added to item 54 comments (original list)
Disclose academic vs industrial use purpose of trial	There is a problem in the literature right now. An over-preponderance of academic trials which do not deliver tools which are suitable for widespread clinical deployment. The papers are written with academic point scoring in mind and do not present a balanced view of the algorithm and its usage. A simple, non-threatening statement would allow the reader to easily class the paper into one of the two categories and read it accordingly.	Partially covered in item 4 (updated list)