

## DECIDE-AI

# Guideline piloting form and results summary

03.09.21

---

DEVELOPERS PERSPECTIVE	2
PEER-REVIEWERS PERSPECTIVE	34

## DEVELOPERS PERSPECTIVE

---

12 full answers (5 grouped in 2 survey) and 2 partial answers (commenting without per item yes/no answers)

Full answers: JEB, AC, (SvdM + RB), HM, MGB, MW, OR, (WYN + DV + TET), YP

Partial answers: FS, LB

### Item 1

*Title and abstract - Identify the study as early clinical evaluation of a decision support system based on artificial intelligence or machine learning, specifying the problem addressed.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

A general comment on the guidelines is that the terminology should be more aligned with that from the Medical Device Regulation 2017/745 in order to be useful for manufacturers making these tools. All manufacturers building medical devices - including Software as a Medical Device (SaMD) using AI/ML - have to comply with the relevant provisions of the MDR if they intend to put their device on the market in the EU. In the MDR the term clinical investigation is used to describe any systematic investigation involving one or more human subjects, undertaken to assess the safety or performance of a device. Clinical evaluation covers the entire process of collecting and evaluating clinical data pertaining to a medical device in order to evidence its safety and effectiveness, for which the clinical investigation (trailing the AI/ML tool in the clinical setting) often provides the bulk of the data. This document, drawn up by the Medical Device Coordination Group, describes some terms proposed for 'pilot clinical investigations' conducted at different phases of development:

[https://ec.europa.eu/health/sites/default/files/md\\_sector/docs/mdcg\\_2021-6\\_en.pdf](https://ec.europa.eu/health/sites/default/files/md_sector/docs/mdcg_2021-6_en.pdf)

It is unclear whether the study design should be mentioned in the title e.g. "Controlled before/after prospective study".

Overall YES. For NLP/chat bot related projects, the difference between early-stage and formative might be blurry. Early stage development might include obtaining performance and soliciting feedback from external collaborators as well.

Providing a list of terms might be helpful in standardizing because people can use the same term with different meanings. An exhaustive list would be impossible to create, but a short list of terms may be possible (e.g. suggest words like 'early-stage', 'formative', etc.)

Perhaps clarify that at the formative stage an algorithm or product would be available for test? Otherwise where's the boundary with early-stage?

Considering your project, would it be possible to report the information necessary for this item?

- Yes 6
- No 3

Comments:

As stated above, difficulty in identifying the actual stage of evaluation.

This is dependent on the type of NLP system. Our project is a general Q&A chatbot which is not necessarily a decision supporting system but plays a role in digital and telehealth. Hence, a broader definition might be required.

Based on your description of the scope of the studies this guideline is targeting, I would probably say something like 'clinical evaluation of --- '

As the study relates to products in service (i.e. beyond stage 2)

## Item 2

*Target problem and population - Describe the target problem and medical condition, including the current standard practice, and the target patient population.*

*Medical indication – Describe the targeted: patient population;*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

suggest medical indication (from MDR) or targeted clinical condition i.s.o. medical condition. Current standard practice could alternatively be worded as current treatment options or state-of-the-art with regards to treatments and/or therapeutic interventions for the condition.

Examples of target “problems” and “medical conditions” would help in the E&E. Would suggest describing target population (e.g. ICU patients) before target problem/outcome (e.g. risk of outcome or likelihood of diagnosis X).

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

Suggestion to not only describe target patient population, but also target for decision support (e.g., nurse, type of physicians)

Intended end users of the tool

Overall YES. However, Natural Language Processing has many utilities (e.g. Medical record analysis and curation, ambient scribe, genomic research analysis) which do not necessarily have a target population.

I would note that there is a wide variation in current practice to difficult to say it is ‘standard’.

## Item 3

*Describe the intended use of the algorithm, its planned integration in the care pathway and the potential impact, including patient outcomes, it intends to achieve.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

Intended use may need some explanation for those not involved in regulation of med tech.

Would be clearer if this was described as an “overview” of how the algorithm was used and integrated into care pathway. Further details to be given in methods.

I would probably expect some of the contents (‘Details about (...) final decisions’) to appear in Methods section and not in detail in the introduction. But I agree in general that these should at least briefly mentioned earlier in manuscript.

Yes, important statement as it sets out boundaries of applicability.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

Assessing patient outcomes with NLP chatbots would be difficult and subjective with no clear guidelines or performance matrix for guidance.

Not sure if I’d be able to comment on ‘the attribution of responsibility/liability for the final decisions’

## Item 4a

*Describe how patients were recruited, stating the inclusion and exclusion criteria at both patient and data level, and how the number of recruited patients was decided.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

‘Consecutive recruitment’: in glossary?

Important to add a rationale for the inclusion and exclusion criteria. This is touched upon in the E&E.

Not all studies will involve active recruitment – for example when using AI to “screen” all adult patients in the hospitals. Might be clearer to refer to participants’ inclusion/exclusion criteria and then report recruitment process (including consent model).

I agree the distinction between patient and data criteria is very important

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

Not completely clear how the recruitment target should be determined, based on comparable studies?

Overall YES. However, there would be difficulties with respect to open-ended (free for all) assessments of conversational Q&A chat bots. While this is more representative of real-life usage, it can result in huge variation in results that could make it hard to judge the required sample size.

## Item 4b

*Describe how users were recruited, stating the inclusion and exclusion criteria, and how the intended number of recruited users was decided.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

Just thinking that from data quality perspective, criteria would be different for patients and users. For patient data you may choose to exclude low quality points, but whatever you collect from the users should all be considered even if they are of low quality and have plans to understand why that is the case (e.g., difficulty in use, lack of interest, etc.)

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

Might be a bit confusing if users are patients as well.

Overall YES. However, similar to 4a, there would be difficulties with respect to open-ended (free for all) assessments of conversational Q&A chat bots. The users are generally the public akin to patients. While this is more representative of real-life usage, it can result in huge variation in results that could make it hard to judge the required sample size.

In my experience low quality user data was still reported and analyzed, possibly with some sensitivity analyses.

## Item 4c

*Describe steps taken to familiarise the users with the algorithm, including any training received prior to the study.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

I'm not sure whether with this you mean describing the inner workings of the algorithm ('opening the black box') to the end users or training them on use of the system/tool/software which hosts the algorithm, with an explanation of how the algorithm works.

Familiarisation could occur within a "wash-in" period during the study, so might be worth mentioning this option specifically in E&E.

This is another important requirement and one that is sometimes buried away without a clear specification.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

However, there might also be no training necessary.

Yes, although it should be noted that some users may have acquired their experience over time in practice rather than specific training sessions. This would be more difficult to report reliably.



## Item 5a

*Briefly describe the algorithm, specifying the version and type of model used. Describe, or provide a direct reference to, the characteristics of the patient population on which the algorithm was trained and its expected performance from development/validation studies.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 7
- No 2

Comments:

Again terminology is a bit confusing ‘development/validation studies’. Validation for manufacturers of SaMD using AI/ML can also mean analytical or technical validation, so looking at the algorithm’s performance only (measured against specified requirements). Suggest to use same terminology from ISO 14155:2020, Annex I describing the different types of clinical investigations at each phase of clinical development. Think UDI, at least in Europe, refers to the SaMD using ML/AI as a whole, not the specific algorithm used.

Not clear what is meant by “type of model” – does this refer to which machine learning/statistical model (e.g. neural network, logistic regression)?

It’d be helpful if you describe what you mean by ‘algorithm’ because it can mean so many different things. The description seems to be referring to publicly available, versioned algorithms (e.g. deep learning models) but not algorithms are like that.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 7
- No 2

Comments:

No version was stated, but it probably should be in hindsight.

Algorithm is trained using developed datasets instead of patient data.

If you go into technical details (such as hyperparameters optimization), this item wouldn’t make much sense).

I can describe the general specifications of the models used (e.g. logistic regression, based on python package version x.x, etc.)

It may be possible, but also could be a challenge to access the required level of detail from a manufacturer.

## Item 5b

*Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, any pre-processing applied and how missing/low-quality data were handled.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

‘including whether data were entered automatically or manually’: into the datasource/EHR from which the input data was collected? Not completely clear.

I would expect that all data used undergoes some form of pre-processing therefore the ‘any’ is redundant. Agree with Siri above that the statement is confusing: is it referring to how the raw patient data was generated (lab data, readings from medical devices, entered manually by a HCP into the EHR etc.) or how the whole data was inputted for training of the algorithm? Nit, but it’s not clear if handling of out-of-distribution data would be best described here or in item 7a

Explanation would be clearer if split into a) how data were acquired (e.g. routinely collected or additional data collection needed for study) b) how specific data items were measured (e.g. blood pressure measurement).

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

‘As different measurement devices and data acquisition techniques can influence the data used as inputs, a description of the acquisition settings should also be considered when appropriate.’: If a large number of variables are used as stored in the EHR, it might be too much to elaborate on each of the measurement devices and data acquisition techniques.

There are no data features input for chatbots. Instead, a description of the how the dataset was developed and expanded could replace the data features.

## Item 5c

*Describe the algorithm outputs and how they were presented to the users (an image may be useful).*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

What is meant by ‘protected identities’ in this context and ‘Any opportunities for users to edit the output’?

I agree, I’m not completely sure what this refers to: ‘the weights of features based on protected identities (e.g. sex, ethnicity)’. Actually would sex be a protected identity? I think it’s encouraged to present sex disaggregated data when reporting on an algorithm although this is probably not something to be presented in the UI.

I would clarify how the “the weights of features based on protected identities” relates to the how outputs are presented to users. For example, if known biases or increased uncertainty based on “protected identities” are presented to users. Would include if/how these biases were assessed in pre-clinical algorithm development and validation.

The time lag between user action and receiving output would be interesting information.

It may also be of interest (depending on the article focus) to indicate the nature of the users’ interaction with the data. For example are they able to manipulate the data-set to see influence of including/omitting certain input variables?

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

Continuing from above, reporting on the interaction mechanisms, such as being able to compare different time periods on a trace would be potentially valuable – but to do this comprehensively would require a lot of description.

## Item 6a

*Describe the settings in which the algorithm was evaluated, including which additional clinical information (i.e. not provided by the algorithm) was accessible to the users to interpret or put into context the output of the algorithm.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 7
- No 2

Comments:

Not completely clear what the additional clinical information would be in our case, all other clinical parameters?

As most SaMD using AI/ML are developed as clinical decision-support tools, I would expect that they would never be used in isolation by the clinician i.e. without access to the patient's data in the EHR, even during a clinical investigation

Was unsure what "visual cues" referred to. Clinicians are likely to have access to many other systems (e.g. radiology, EPR), as well as the UI where the algorithm results are presented. I would also report whether the study recorded how often these systems were accessed by clinicians when assessing the algorithm results.

This is an admirable goal, but in practice could require a lot of resource and be difficult to report, given the wide range of clinical and contextual factors that influence decision-making.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 6
- No 3

Comments:

Would result in an elaborate description of the clinical workflow and factors influencing the decision for diagnostics/treatment.

This would be difficult to ascertain if it was conducted online or via mobile platforms. The required reporting can only be achieved at a controlled research site which is not a realistic representation of real-life usage.

However, it might be almost impossible to list all the information a user could access in the Electronic Health Records of an hospital.

Difficult given the broad area of clinical and contextual factors that influence a decision (not just the output of the CTG machine).

## Item 6b

*Describe the clinical workflow/pathway in which the algorithm was evaluated, the timing of its use, how the final decision was reached and by whom.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

Glossary: concurrent vs. second reading

It is important to state the significance of information provided by the algorithm to the healthcare decision (from IMDRF/SaMD WG/N12FINAL:2014) : to treat or to diagnose; to drive clinical management or to inform clinical management. Suggest to change 'how the final decision was reached' to 'describe the significance of the algorithm's output to the clinical decision-making process'.

The E&E currently is more a justification of including this item, rather than focusing on key required reporting elements. The E&E could be clarified with more examples of how the clinical pathway and decision-making process are described.

I think the detailed decision-making in the article is too much information. As physician, we now that a lot of different aspects influence the patient's final decision and we don't need to know details about that, only how many followed the algorithm suggestion or not in real life.

As comment for 6a) One could endeavour to report the most important influencing aspects, but to do so comprehensively would require a lot of resource and may be impracticable.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 7
- No 2

Comments:

See above. We are planning to include 600 patients in our DECIDE-AI research, it would be impossible to discuss in such detail the final decision.

There might not be a final decision involved for Q&A related chatbots, unless measured by matrices e.g. % conversion from chatbot to call-in inquiries.

Difficult given the broad area of clinical and contextual factors.

## Item 7a

*Provide a description of how significant errors/malfunctions were defined and identified.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

How elaborate should this be, a full potential hazard analysis?

I would say there are more than 3 categories of errors as listed in the E&E (e.g. data)

E&E would benefit from more examples. Do algorithm errors refer to its performance or to situations where the algorithm software failed? What would count as a “user error”?

Not sure what is meant by algorithm error given that typically they produce a probability/confidence mediated output. If an AI device returns a 90% probability for absence of a signal, but on subsequent inspection a signal was present, that is failed output, but not necessarily an error.

I would separate out error and malfunction. The former suggests a validation issue (AI does not meet purpose), the latter suggests a reliability issue (failed component).

In medical device testing standards the term “use error” rather than user error is used to reflect the problems may lie in the interface rather than the human. This same principle could apply to poor presentation of algorithm outputs.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

As above.

## Item 7b

*Describe how any risks to patient safety or observed instances of harm were identified, analysed, and minimised.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

‘This builds on Item 8a:’ Should be item 7a?

Note to E&E and question: SaMD using AI/ML manufacturers are expected to have conducted risk management (a process to identify hazards, to estimate and evaluate the associated risks, to control these risks and to monitor the effectiveness of these controls) before performing any clinical investigation. ISO standard 14971:2019 and the accompanying document ISO/TR 24971:2020 are useful references for this. The following could be added for clarity: ‘Describe your risk management process, ...’ In addition, as part of their clinical investigation plan (see ISO 14155:2020) manufacturers have to describe the clinical benefits and risks of the ‘investigational device’, including any anticipated adverse device effects and risks associated with participation in the clinical investigation.

Would specifically report whether a pre-study risk assessment was undertaken and how.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

Not applicable for a general Q&A chatbot. This might be more relevant to symptom-based triaging AI chatbots.

## Item 8

*Describe the human factors tools, methods or frameworks used, the use cases considered, and the users involved.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

Not completely clear what is meant by ‘*human factors tools*’

Would add to the E&E that all SaMD using AI/ML manufacturers have to comply with the requirements from IEC 62366-1:2015/AMD1:2020 (Application of Usability Engineering to Medical Devices)

I am not sure everyone will be familiar with “human factors”. It might be worth expanding to include implementation science and user interface methodology.

It’s good that you reference the medical device usability standards. In practice these are often narrowly applied in lab settings with limited scenarios, so I like the pointers to contextual factors.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

I don’t think we would use – but again, perhaps we should have.

In particular I would also use methods to elicit the judgement and rationale used by users to accept/reject and interpret the AI output.



## Item 9

*Describe whether specific methodology was utilized toward an ethics-related goal (such as algorithmic fairness) and its rationale.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

I would split this into evaluation of algorithm “fairness”/bias undertaken pre-, during- and at the end of the study.

Also if there is any plans for continued monitoring of fairness/ethics related goals

This is an area I would not be confident reporting upon. However, the E&E provides references and with consultation of a specialist this is acceptable.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 7
- No 2

Comments:

Not really applicable for intraoperative videos.

We have already bias in our datasets to develop our algorithms. So, for now, it would be impossible to promote fair risk assessment. We still have to gather more diverse data to be fair.

I would need to seek advice on how to address potential biases. Sensitivity testing required perhaps?

## Item 10a

*Describe the baseline characteristics of the patients included in the study, and report on input data availability.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

this is quite similar to item 4a

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

No patient data is required and no input data is selected. Only training and testing dataset is required.

## Item 10b

*Describe the baseline characteristics of the users included in the study.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

Would expand E&E to include standard descriptors (e.g., age, gender, profession, seniority). It is possible that there may be few clinical “users”, so descriptors should preserve anonymity.

Baseline characteristics of the user seems very vague (type of healthcare professional, sex, age, years of experience, digital literacy, the list can go on)

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

## Item 11a

*Report on the user exposure to the algorithm, on the number of instances the algorithm was used and on the users' adherence to the intended implementation.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

'because such deviations could bias the results of the study.': Not sure if this also applies to assistive decision support, as physicians can choose to deviate from the suggested action which also will occur after implementation.

See my earlier comment about the significance of the information provided by the algorithm/SaMD on the healthcare decision.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 7
- No 2

Comments:

Not sure this would be easily recordable.

This will be possible in a small controlled group. However, chatbots are intended for large scale use and depending on the backend design, the details required might not always be available.

## Item 11b

*Report any significant changes to the clinical workflow or patient pathway caused by the algorithm.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

Not exactly clear what should be reported in this section, should it be quantitative reporting of changes in time spent on different tasks for the intervention arm compared to the non-intervention arm?

The E&E para could perhaps help authors structure how this information is recorded better (e.g. by referencing Item 3 and how real world implementation differed from anticipated use)

Would also report how “changes” were detected/recorded. Would separate out changes to clinical workflow (e.g. additional data recorded on “live” system) versus patient interventions. In E&E, I would specifically mention monitoring patient interventions to relate these to the outcome.

One challenge may be the inconsistent use of the AI functions (for example used by some individuals or wards and not others, or used depending on issues such as staffing). Good to endeavour to report these workflow changes though.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

Might be difficult to measure.

Yes when the project progresses to a later stage development.

## Item 12

*Report any changes made to the algorithm or its hardware platform during the study. Report the timing of these modifications, the rationale for them, and the change in outcomes observed after each of them.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

Very clear.

Just as an aside, it would be very unlikely that during a formal clinical investigation any significant changes would be allowed to the algorithm, unless they are essential for safety reasons.

May want to consider adding reporting of whether timing of modifications was accounted for in the analyses.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

Some changes can be made outside of the study but still affect the study (e.g. update to EHR system)

## Item 13

*Report on the user agreement with the algorithm. Describe any instances of and reasons for user variation from the algorithm's recommendations and, if applicable, users changing their mind based on the algorithm recommendations.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

Should the user report their opinion for each time the CDSS is used?

It might be worth introducing this in the Methods section, if user agreement should be a key outcome of the study.

Just occurred to me while reading this section – do you only consider algorithms that output explicit recommendations? Some CDSS can be more focused on providing additional insights rather than prescribing actions.

I think there is an overall scoping point for the guidelines – do they only include “decision support” systems? AI can be applied to problems/tasks with minimal human intervention once running.

Even if we assume only decision support systems I've answered no as I think the table of potential outcomes is interesting but oversimplifies. For example the AI system may automatically action something based on pre-set parameters with the human able to veto; this is different to a system in which a recommendation is presented (perhaps with confidence indicators) and the human must initiate action or ignore. In the table these both may be listed as “human diverging from AI recommendation” but the implications (in terms of ethics, performance and benchmarks) are quite different. Administration of drugs may be one example.

In summary, in medical domain the range of actions to outcomes do not have clear “correct/incorrect” relationships.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 6
- No 3

Comments:

Not sure this would be easily recordable.

This is more relevant to outcome/symptom-based triaging chatbots. General Q&A chatbots differ in intent, with users/patients looking for information which may or may not produce an actionable event. Users/patients opinion regarding the relevance of the answer would be reported.

As mentioned earlier determining “correct” versus “incorrect” outcomes could be difficult to measure in field-based work. For the CTG case I have in mind it would be very difficult to state that a AI output was correct/incorrect.



## Item 14a

*List any significant errors/malfunctions related to: algorithm recommendations, supporting software/hardware, or users. Include details of: (i) rate of occurrence, (ii) apparent causes, (iii) whether they could be corrected, and (iv) any significant potential impacts on patient care.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

E&E would perhaps benefit from standard metrics (e.g., number of hours system was non-functional).

Yes, but as hinted at in the E&E, comprehensive testing should have occurred before this phase and should include automated testing methods.

On matters of judgement such as AI might provide (listed as error) it could be difficult at this stage of development to say if it is a good or poor recommendation. Independent recommendations would be needed for comparison.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

To give the example of a CTG trace: an automated CTG recommendation can difficult to prove/disprove as correct because there is not a clean distinction between 'normal' and problematic trace features – the delta is low, and the true state of the fetus is unknown.

## Item 14b

*Report on any risks to patient safety or observed instances of harm (including indirect harm) identified during the study.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

How does this differ from Item 7b?

Would relate this to 7b) in E&E.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

## Item 15a

*Report on the usability evaluation, according to recognised standards or frameworks.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

Very clear description. Although I think usability testing should be done before entering the clinical trial phase.

Yes, I would expect that usability testing with the UI has already been completed before any investigation in order to identify and reduce user errors.

Would relate to item 8, as the metrics used for usability evaluation should be decided a priori. Would add comparison to pre-implementation usability evaluations, if relevant.

What recognized standards or frameworks? This is not clear to me.

From my own experience, I'd want to see more on what's expected from the Usability evaluation section. There are some useful flow charts in BS EN 62366-1:2015+A1:2020 to reference as examples and illustrate how that dovetails with the whole process.

Good to see the standards referenced. You could make reference to the opportunities and limitations of lab-based usability testing versus in-situ testing.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

[yes] We did not plan to use ISO standards, but some relevant evaluation.

[no => no explanation, see "What recognized standards or frameworks?"]

## Item 15b

*Report on the user learning curves evaluation.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8
- No 1

Comments:

I was unclear how to assess “learning curves”. Would suggest adding to Method section with examples.

User learning curve may not be known to everybody.

Could provide some example of proxy measures as I wasn’t clear whether this was task-related or not.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 8
- No 1

Comments:

Yes, although it would need some discussion to decide on which learning curve metrics to use for our project.

Chatbots, by the very fact that they are a User Interface, places heavy emphasis on User Experience. Hence, most chatbots naturally place great emphasis on minimizing the learning curve. As a result, nearly all chatbots are by nature easy to use and comprehend. Hence, value of reporting might be limited.

## Item 16

*Discuss whether the results obtained support the intended use of the algorithm in clinical settings.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

Would specifically mention reporting limitations of study in the E&E or refer to “good research practice” section.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

## Item 17

*Discuss what the results suggest about the safety profile of the algorithm. Discuss the observed errors/malfunctions and instances of harm, their implications for patients and whether/how they can be mitigated.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 9
- No 0

Comments:

Not completely clear what is meant by '*safety profile*'

It's fine but again this seems like a lot of overlap.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 9
- No 0

Comments:

## Item 18

*Disclose if and how data and relevant code are available.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 8.5
- No 0.5

Comments:

data is unclear - used to train the algorithm, patient data used during the investigation, clinical and technical performance data generated during the investigation?

E&E would benefit from examples of types of data (e.g., summary outcome data versus patient-level) that could or could not be shared. Would also emphasise that a machine learning/statistical model could be shared, without sharing the code used to develop/validate it.

Considering your project, would it be possible to report the information necessary for this item?

- Yes 7
- No 2

Comments:

At this moment, we are discussing with our Funder if the code could be available to third parties.

In practice the manufacturers would be unlikely to reveal their code.

## Good research practice items

See “*Final\_good\_research\_practice\_list.docx*”

Does the division between the AI-specific and good research practice item list makes sense to you? Is this division easy to understand?

The division is clear as the good research practice item list applies to all fields of research and is not limited to research on AI decision support tools. The relevance of adding these guidelines to the DECIDE-AI ones is not completely clear for me, as in my opinion following the AI-specific item list would most likely result in a trial that follows the items as stated in the good research practice list. After reading the E&E list, the addition of these items and their relevance for research on AI decision support tools becomes more clear.

Yes; and the materials are comprehensive.

Yes

Yes, but it would be helpful to link each AI-specific item to a good research practice item.

It does make sense. I think it is helpful.

Yes to both.

The division is easy to understand. Item 4a and 4b, 10a and 10b from the AI list could be placed in the good research practice list as they are not really specific to AI.

The list makes sense to me. I see how they are divided (AI specific vs. general), but possibly because it's something I'm familiar with and also something I've thought about for a while.

Do you have any comments on the wording or E&E paragraphs of the items in the good research practice list? (Please specify the item numbers when applicable.)

None

No

No, it is fine.

They are well structured and succinct.

None



## Used terminology

After reading through the guidelines, which of the following terms do you think would be the clearest and most appropriate to describe the evaluated system: “algorithm”, “decision support system”, “AI system”, “AI-based clinical decision support system (AI-CDSS)”, “decision support tool”, “AI tool”?

2 votes for “AI-based clinical decision support system (AI-CDSS)”

AI-based clinical decision support system (AI-CDSS)

AI-based clinical decision support system (AI-CDSS)

Decision support tool would be the most appropriate description. This would cover algorithms embedded in existing “systems” (e.g., EPR, radiology software).

I think “algorithm” is the simplest and most understandable term for everyone.

AI system.

AI-CDSS

Do you have any other suggestions?

I think that it will be impossible to describe all items in this pilot in only one paper, even with the help of supplementary files. Perhaps it should be divided in two types of articles: one focusing on human aspects of the use of AI algorithms and another focusing on the technical aspects of the development, metrics, missing data, hardware, etc, because even for the reader, let alone the authors, it is too much information.

Deep learning system (if Deep learning is used); AI model;

**Thank you very much for your participation. Your feedback is much appreciated and will play an important role in making the DECIDE-AI guidelines more useable and impactful.**

## PEER-REVIEWERS PERSPECTIVE

---

2 full answers

Full answers: OA, LL

### Item 1

*Title and abstract - Identify the study as early clinical evaluation of a decision support system based on artificial intelligence or machine learning, specifying the problem addressed.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes                      2
- No                        0

Comments:

The thing that slightly confused me was the list of potential terms. After all, you include both phase 2 and stage 2. I would perhaps make the point that you are not advocating these particular titles.

A general point please. You seem to mix Roman and Latin numerals. So in document 4 and 5, the items are in Roman numerals. In this document you are using Latin numerals. I think they are the same thing but I am rather confused. In document 4 and 5, the items have titles but they do not appear in document 3. And in doc 4, you do not include the sub-items that are in doc 3. Are they the same things?

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes                      2
- No                        0

Comments:

## Item 2a

*Target problem and population - Describe the targeted medical condition and problem, including the current standard practice, and the intended patient population.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Very clear and appropriate

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 2b

*Describe the intended users of the AI system, its planned integration in the care pathway, and the potential impact, including patient outcomes, it intends to achieve.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Very clear and appropriate

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 3a

*Describe how patients were recruited, stating the inclusion and exclusion criteria at both patient and data level, and how the number of recruited patients was decided.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes                      2
- No                        0

Comments:

I like the comment that there may not be a statistical model at the beginning. This is important as many journals may be looking for this information. It might be valuable to highlight this even more. We often spend ages trying to manufacture a statistical output that will reassure reviewers. I would suggest that authors are required to highlight specifically that they think it is inappropriate for this procedure and they are therefore not doing it.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes                      2
- No                        0

Comments:

Again, I would highlight the fact that statistical sample size calculations might not always be appropriate, although it is perhaps a good idea for authors to state specifically that they think it is inappropriate for this procedure and they are therefore not doing it.

## Item 3b

*Describe how users were recruited, stating the inclusion and exclusion criteria, and how the intended number of recruited users was decided.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes                2
- No                0

Comments:

I strongly support this. It is important. Similar comments about statistical issues as in 3a.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes                2
- No                0

Comments:

## Item 3c

*Describe steps taken to familiarise the users with the AI system, including any training received prior to the study.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

I just wonder about the comment 'ideally a training protocol and the training materials... should be included in ... appendices. There is often going to be a commercial conflict of interest. Is there something that needs to be considered here about this (I see that this is the last section of the good research practice list but it might be that more advice is needed about this.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 4a

*Briefly describe the AI system, specifying its version and the type of underlying algorithm used. Describe, or provide a direct reference to, the characteristics of the patient population on which the algorithm was trained and its expected performance from preclinical development/validation studies.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 1
- No 1

Comments:

I don't know what Standardised AI system facts labels are. If you want to leave this comment in, rather than just referring to a paper, put in a few words to explain the concept.

It is highly likely that some researchers may have already published their in-silico/pre-clinical study, which describes the algorithm and associated training data, hence a reference might suffice.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:



## Item 4b

*Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, the pre-processing applied and how missing/low-quality data were handled.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes                2
- No                0

Comments:

From experience this might be tricky. It is a good idea but not always possible to define precisely all the parameters suggested here. But it is a good idea. If journals demand that people must add in all the information I can envisage some papers getting into trouble. I am not sure how to make the statements less onerous however 😊

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes                2
- No                0

Comments:

But see comment above

## Item 4c

*Describe the AI system outputs and how they were presented to the users (an image may be useful).*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

I can envisage that this might be overkill. So if an AI has been developed already and the HCI is fixed, all that is needed is a reference to a previous paper. I would suggest adding this as a comment. This is particularly relevant when the focus of the paper is not the AI HCI but rather the outputs.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

But see above.

## Item 5a

*Describe the settings in which the AI system was evaluated.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 5b

*Describe the clinical workflow/care pathway in which the AI system was evaluated, the timing of its use, how the final decision was reached and by whom.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes                      2
- No                        0

Comments:

In item 1, you gave some suggested terms to use. It might be useful to do so here eg detection, characterization, diagnosis, decision support etc. It would be quite useful to offer options which would help standardize reporting

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes                      2
- No                        0

Comments:

## Item 6a

*Provide a description of how significant errors/malfunctions were defined and identified.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

As I read this, I wonder about whether there should be a mandatory checklist to include. You would simply need to add an extra column in document 2 that the authors can tick to confirm that they have included this item

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

It might be worth making it easier for reviewers by changing the supporting document so that each of the bullet points I to iii are on separate lines.

## Item 6b

*Describe how any risks to patient safety or observed instances of harm were identified, analysed, and minimised.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 1
- No 1

Comments:

‘Ideally...’ What happens if the authors do not report this. Is it enough just to say that the system fulfils ISO14971 etc? I have no idea about this sort of detail but if it is important I would ask the AI company we work with. It does not sound particularly meaningful to me

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 1
- No 1

Comments:

I have no knowledge of these ISO standards (should I have?!)

## Item 7

*Describe the human factors tools, methods or frameworks used, the use cases considered, and the users involved.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 1
- No 1

Comments:

Is this mandatory? It does not seem to me that this is always necessary.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 1
- No 1

Comments:

Again, how many different HCI parameters are you suggesting that are included? If this is important, I suggest you list out clearly which parameters must be included and which are optional.

## Item 8

*Describe whether specific methodologies were utilized toward an ethics-related goal (such as algorithmic fairness) and their rationale.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes                      2
- No                        0

Comments:

I have never considered this in an AI study of detection of lesions during endoscopy. Presumably the idea is that people should simply put in a line stating that no specific methodology was used towards any ethics-related goals. It might be helpful to suggest a standard wording to use in that situation.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes                      1.5
- No                        0.5

Comments:

As above, this is not always relevant but it appears that you are asking for it to always be included.



## Item 9a

*Describe the baseline characteristics of the patients included in the study, and report on input data availability.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 1.5
- No 0.5

Comments:

Study population characteristics makes sense. For some studies, input data availability makes sense. But in slightly more advanced studies, input data availability might be just too much detail and not central to the running of the study. I am not sure I like this requirement.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 9b

*Describe the baseline characteristics of the users included in the study.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 10a

*Report on the user exposure to the AI system, on the number of instances the AI system was used, and on the users' adherence to the intended implementation.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

It is clear but it may not always be relevant. My question remains whether all these sections are 'required' or if some can be optional and used only if appropriate

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 10b

*Report any significant changes to the clinical workflow or care pathway caused by the AI system.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Same as last section. This is perfectly reasonable for certain studies but not for all. Should be optional.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 11

*Report any changes made to the AI system or its hardware platform during the study. Report the timing of these modifications, the rationale for them, and the change in outcomes observed after each of them.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 1.5
- No 0.5

Comments:

I am not sure that I agree with this part: “provided that there is no attempt to make an overall summative conclusion about device effectiveness”

As AI technology is developing so fast, even in a randomized controlled trial of an AI that takes a year to perform, the system could be much better at the end than at the beginning. As long as the randomization was done in a way to ensure that the control procedures were being done at the same time as the AI ones, incremental changes to an AI system would not necessarily be unreasonable.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

Useful for some studies, but see comments above

## Item 12

*Report on the user agreement with the AI system. Describe any instances of and reasons for user variation from the AI system's recommendations and, if applicable, users changing their mind based on the AI system recommendations.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 1
- No 1

Comments:

The table is very well designed but it is not referenced in the text. Do you want such a table included in every paper or is it just placed there to help people understand the different scenarios? Need to rethink this slightly.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 1
- No 1

Comments:

As above

## Item 13a

*List any significant errors/malfunctions related to: AI system recommendations, supporting software/hardware, or users. Include details of: (i) rate of occurrence, (ii) apparent causes, (iii) whether they could be corrected, and (iv) any significant potential impacts on patient care.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes            2
- No            0

Comments:

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes            2
- No            0

Comments:

## Item 13b

*Report on any risks to patient safety or observed instances of harm (including indirect harm) identified during the study.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:



## Item 14a

*Report on the usability evaluation, according to recognised standards or frameworks.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

The list of standards feels very UK centric. There is no mention of any USA or Asian standards, although you mention an International standard set. Are there other examples to add here?

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 14b

*Report on the user learning curves evaluation.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 1.5
- No 0.5

Comments:

Again, this will only be relevant in a sub set of studies. This should be an optional item in my opinion.

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 15

*Discuss whether the results obtained support the intended use of the AI system in clinical settings.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 16

*Discuss what the results suggest about the safety profile of the AI system. Discuss the observed errors/malfunctions and instances of harm, their implications for patients and whether/how they can be mitigated.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Item 17

*Disclose if and how data and relevant code are available.*

After reading the item and E&E paragraph, is it clear what should be reported in this section?

- Yes 2
- No 0

Comments:

Would the information reported under this item be useful to you in order to peer-review the study?

- Yes 2
- No 0

Comments:

## Good research practice items

*See “Final\_good\_research\_practice\_list.docx”*

Does the division between the AI-specific and good research practice item list makes sense to you? Is this division easy to understand?

I would simply say that the AI specific item list is there to supplement the good research practice item list. There is clearly some overlap but that does not matter as the new list is to help authors specifically in this field.

In the AI specific list and the good research practice item list, they both currently have ‘title and abstract’ under the topic of title and abstract.

However the AI specific list seems to be relevant specifically to the title, whilst the good research item list seem to be relevant specifically to the abstract.

Do you have any comments on the wording or E&E paragraphs of the items in the good research practice list? (Please specify the item numbers when applicable.)

My only question is whether the 2 lists could be merged. If I am writing a paper, I want to look at a single checklist to ensure I have included everything. Having to look at 2 checklists, which in any case overlap, could be rather tiresome and repetitive, as well as confusing.

**Thank you very much for your participation. Your feedback is much appreciated and will play an important role in making the DECIDE-AI guidelines more useable and impactful.**