

# DECIDE-AI - Delphi 2nd round

New reporting guidelines for the Developmental and Exploratory Clinical Investigation of DEcision support systems driven by Artificial Intelligence (DECIDE-AI).

## Annex V-10

---

PARTICIPANT INFORMATION SHEET  
Delphi round 1 and 2  
Version 1 - 27.11.20

Developmental and Exploratory Clinical Investigation of DEcision support systems driven by Artificial Intelligence (DECIDE-AI): development of new reporting guidelines through a Delphi process.  
CUREC Approval Reference: R73712/RE002

We'd like to invite you to take part in our research study. Before you decide, it is important that you understand why the research is being done and what it would involve for you. Please take time to read this information, and discuss it with others if you wish. If there is anything that is not clear, or if you would like more information, please don't hesitate to contact us.

**What is the purpose of the study?**

This study aims to develop new reporting guidelines to make the early clinical evaluation of AI driven algorithms more consistent, comprehensive and reproducible. A robust evaluation at this stage where algorithms are first used by clinicians is indispensable to bridge algorithm development to large-scale clinical trials. To this end, DECIDE-AI ambition is to improve the evaluation reporting along four main axes:

i) the algorithm performance when first used with humans in small-scale, representative clinical conditions, ii) the safety profile of the algorithm prior to its wider-scale utilisation, iii) the human factors (ergonomic) evaluation of the algorithm, and iv) the preparatory steps towards large-scale clinical trials. A more detailed description of the project can be found in the document "DECIDE-AI project presentation".

**Why have I been invited?**

You have been invited on account of your expertise and experience in the field of artificial intelligence and/or evaluation of clinical intervention. We believe that your expert opinion will contribute to develop comprehensive and well-informed reporting standards.

**Do I have to take part?**

No. Please note that participation is voluntary. If you do decide to take part, you may withdraw consent at any point for any reason, and without any adverse consequences or penalty. You can exit the survey at any time before submitting your answers by closing the browser. If you need to interrupt your session but wish to continue participating, you can choose the option to save your progress and return later. If you do not wish to answer a question, you can simply fill the answer field with "NA" or choose the "I don't know" option.

**What will happen to me if I decide to take part?**

The research project is based on the published Delphi methodology. The main goal of a Delphi process is to reach consensus between experts through several rounds of feedback. The results of each round are presented during the following round to inform participants' decision and guide them toward consensus. The present Delphi adaptation will include two general rounds, followed by a consensus meeting (for a subset of participants). The present invitation and information sheet refer to the two general rounds only. Consent will be asked before each round through an online form.

**Round 1 (online survey):** you will be asked to answer open-ended questions about what you think should be reported when evaluating an artificial intelligence based algorithm for the first time in clinical settings. You will then be asked to score, on a 1-9 scale and according to their importance, a list of provisory items developed by the research team and reviewed by the DECIDE-AI steering group. You will also have the opportunity to add comments, propose new items and recommend additional experts to take part in the Delphi. This round should take you between 45 and 60 minutes, depending on the extent of your free text inputs. You can interrupt your session at any time and resume later.

**Round 2 (online survey):** you will be asked to score, on a 1-9 scale and according to their importance, a modified list of reporting items, updated based on the first-round results and feedback. You will again have the opportunity to provide open comments. This round should take you between 30 and 45 minutes, depending on the extent of your free text inputs. You can interrupt your session at any time and resume later.

**Are there any possible risks from taking part?**

This study is considered to be at very low risk of physical or psychological harm. Issues related to data breaches or loss of confidentiality cannot be totally excluded. However, the research team took this aspect very seriously and designed a data handling strategy to minimize this risk.

#### How will my data be used?

The data we will collect that could identify you will be: your name, your affiliation, your main professional geographical location, your stakeholder group, your experience with AI/clinical evaluation and your professional email address. Your answers will be dissociated from these data using the REDCap software and data analysed in a de-identified manner. We will take all reasonable measures to ensure that your answers remain confidential. Your personal data and answers will be retained in the REDCap DECIDE-AI project database. REDCap is a secure web application, developed by a multi-institutional consortium initiated at Vanderbilt University. Access to the REDCap DECIDE-AI project database is password protected and for authorized users only. Additionally, your personal data and de-identified questionnaires' answers will be stored on a password-protected university network drive. Any linkage between your personal data and questionnaires' answers as well as your professional email addresses will be deleted one month after the publication of the study results. De-identified questionnaires' answers and consent records will be stored for at least three years on a password-protected university network drive. Your name, affiliation, stakeholder group, experience with AI/clinical evaluation and main professional geographical location will remain part of the study outputs (see below).

#### Who will have access to my data?

The University of Oxford and REDCap are the data controller with respect to your personal data and, as such will determine how your personal data is used in the study. The University will process your personal data for the purpose of the research outlined above. Research is a task that we perform in the public interest. Further information about your rights with respect to your personal data is available from <https://compliance.admin.ox.ac.uk/individual-rights>. The data you provide will be de-identified before it is shared outside the research team. As required by the Delphi methodology, anonymous (and often aggregated) answers will be presented to the other participants during the second round, to members of the steering group and to members of the consensus group. Anonymous quotes and anonymous aggregated answers may be used in academic publications.

#### Will my taking part in the study be kept confidential?

In your capacity as invited expert, and for the transparency of the process, your name and affiliation will be disclosed in the final publication. However, your answers will be kept confidential and no linkage between your personal information and answers will be shared outside the research team.

#### Will I receive compensation for taking part?

There will be no financial or in-kind compensation for taking part.

#### What will happen if I don't want to carry on with the study?

You may withdraw consent at any point for any reason, and without any adverse consequences or penalty. If you withdraw your consent up to ten working days after the end of a Delphi round, all data collected during this round will be deleted. Otherwise, your personal information will be deleted but the rest of your answers will remain part of the analysis.

To withdraw consent after completing one or both of the surveys, please contact the primary researcher or Principal Investigator (details below).

#### What will happen to the results of this study?

The results of this study will be published in a peer-reviewed journal and advertised through social media platforms. Participants will have full access to the results on request to the research team. This project will be written up for a DPhil degree.

#### Who is organising and funding the study?

The Principal Researcher is Baptiste Vasey and the Principal Investigator is Prof Peter McCulloch, who are affiliated to the Nuffield Department of Surgical Sciences at the University of Oxford. The project is carried out in collaboration with the DECIDE-AI Steering Group

No specific founding was acquired for the project. Funding is available if required from the IDEAL Collaboration research group general funds. BV is supported by a Berrow Foundation Lord Florey scholarship.

#### Who has reviewed this study?

This project has been reviewed by, and received ethics clearance through, the University of Oxford Central University Research Ethics Committee [reference number].

#### Who do I contact if I have a concern or I wish to complain?

If you have a concern about any aspect of this study, please speak to Baptiste Vasey ([baptiste.vasey@nds.ox.ac.uk](mailto:baptiste.vasey@nds.ox.ac.uk)) or their supervisor Prof Peter McCulloch ([peter.mcculloch@nds.ox.ac.uk](mailto:peter.mcculloch@nds.ox.ac.uk)/+44 (0)1865 740870), and we will do our best to answer your query. We will acknowledge your concern within 10 working days and give you an indication of how it will be dealt with. If you remain unhappy or wish to make a formal complaint, please contact the Chair of the Research Ethics Committee at the University of Oxford who will seek to resolve the matter as soon as possible:

Medical Sciences Interdivisional Research Ethics Committee;

Email: [ethics@medsci.ox.ac.uk](mailto:ethics@medsci.ox.ac.uk);

Address: Research Services, University of Oxford, Wellington Square, Oxford OX1 2JD

---

CONSENT

---

First name:

---

Middle name:

---

Last name:

---

---

Please note that you may only participate in this survey if you are 18 years of age or over.

- ☐ Yes  
☐ No

I certify that I am 18 years of age or over:

---

I have read the information above and agree to participate with the understanding that the data (including any personal data) I submit will be processed accordingly:

- ☐ Yes, I agree to take part  
☐ No, I don't wish to take part

---

**ROUND 1 RESULTS**

---

**OVERVIEW** Thank you once again for your valuable insights and feedback in the Round 1 questionnaire. We received 121 completed questionnaires, with over 46000 words of free text answers to the open-ended questions, 6480 item scores, 312 item comments and 64 propositions of new item to analyse. In the document below, you will find:

an overview of the Round 1 analysis (executive summary) the detail of the Round 1 results (detailed scores and comments for each item of the original item list, thematic analysis based on the answers to the open-ended questions and summary of the new items proposed by the participants) the resulting updated item list. We would like to encourage you to read the executive summary and get familiar with the overall structure of the updated list as this will inform your answers for the current questionnaire.

---

Round 1 analysis - executive summary

[Attachment: "Round1\_executive\_summary.pdf"]

---

Round 1 results - per item scores and comments

[Attachment: "Round1\_per\_item\_results.pdf"]

---

Round 1 results - thematic analysis

[Attachment: "Round1\_thematic\_analysis.pdf"]

---

Round 1 results - proposed new items (participants)

[Attachment: "Round1\_proposed\_new\_items.pdf"]

---

Round 2 - updated item list

[Attachment: "Round2\_updated\_item\_list.pdf"]

---

**DECIDE-AI SCOPE**

---

After analysing the general comments in Round 1 and before starting Round 2, the research team would like to give some clarification about the intended scope of DECIDE-AI.

DECIDE-AI focuses on the early-stage clinical evaluation of AI-based decision support systems, independently of their study design, which might vary depending on the intended purpose of the algorithm evaluated. In other words, DECIDE-AI are guidelines targeted towards a specific development stage rather than study design, similar to the IDEAL reporting guidelines. Some of the assumptions made at this stage of evaluation are that the preclinical development (initial training and testing, preclinical safety and usability testing) of the algorithm has been completed, adequately reported and has shown results supporting a translation to clinical evaluation.

Although most of the DECIDE-AI items will likely also apply to "silent mode" or shadow evaluation (i.e. during which the decision support is evaluated in parallel to the clinical workflow), the main focus of the guidelines is on real life use of the algorithm at small scale, in which the assisted decisions taken have actual impact on patient care.

Many studies within the scope of DECIDE-AI will be testing medical devices as part of evidence generation for regulatory approval. The focus of DECIDE-AI is on scientific reporting and, although this should align as much as possible with regulatory interest, the guidelines are by no means a substitute to other regulatory reporting or research governance requirements.

Usability testing and safety assessment are important parts of the preclinical evaluation of medical devices. The reporting for DECIDE-AI is again no substitute for these evaluations. However, as emphasised by our preliminary literature search and participant comments, many aspects related to usability and safety can only be approximated during preclinical studies and need validation under real world conditions. Moreover, use in actual clinical settings might highlight new issues not previously considered that are relevant for both large scale testing and deployment. These aspects fall within the focus of DECIDE-AI.

If you have major comments or reservations about the overall scope of DECIDE-AI and have practical propositions about how this can be addressed, we would like to give you the opportunity to share your arguments and recommendations for improvement with the Consensus Group. Please send us a maximum 500 words (1 page) text before May 19th 2021 and we will include it in the consensus meeting preparatory material.

---

**ROUND 2**

---

**OVERVIEW** In this second round you are invited to score and comment on an updated item list. This update consists of a reorganisation (merge and/or split), addition, deletion and rewording of items, based on the feedback received from Round 1. This approach was chosen to integrate participant opinions as much as possible, and give you the opportunity to now comment in Round 2 on an item list that is hopefully closer to the final reporting guidelines.

The updated item list will be presented to the Consensus Group as you see it. Therefore, the objectives of this second round of Delphi are:

to highlight the items for which there is consensus; to provide actionable indication and comments to the Consensus Group, in order for them to select the final items to be included in the DECIDE-AI guidelines. When updating the item list, the research team inclined deliberately towards including too many items proposed by participants, rather than excluding items before again seeking the opinion of the Delphi's experts. Therefore, the current item list is too long and covers more aspects than researchers in the field can reasonably be expected to report in a single publication. During this second round, the research team would appreciate your guidance on where to shorten it.

---

**INSTRUCTIONS** Round 2 is broken down into 2 main tasks:

Score each item of the updated item list on a 9 point scale. When doing so, please keep in mind the following score legends: 1-3: not important; 4-6: important but not critical; 7-9: important and critical. In other words, 'not important' items are unnecessary; 'important but not critical' items could be included but are not indispensable; 'important and critical' items form the minimum reporting standards. These scores will build the basis of the Consensus Group discussions and decisions. Comment, with justification, on any specific items you feel need particular attention during the consensus meeting. This could be to back your recommendation to include/exclude this item, propose to merge/split it or suggest a change in wording. These arguments will be presented to the Consensus Group. As any consensus process, the DECIDE-AI Delphi is susceptible to some degree of bias. Although the research team has tried its best to mitigate them, not all could be controlled. When answering Round 2, please be mindful of:

The anchoring bias inherent in the study and questionnaire design. When making judgments, we tend to give more importance to information already provided. Your own framing bias. We all come with a certain opinion on the subject influenced by our experience and expertise. To reach consensus, it is important to consider other stakeholder groups' perspectives as well. There is a total of 53 items in this round.

Technical note on REDCap: If you need to interrupt your session, please use the "Save and Return Later" button (at the bottom of each page). If you close the tab or window directly, you will have to start over from the beginning. A few participants experienced technical issues with REDCap during Round 1. If when moving to the next section or using the "Save and return later" button, you are logged out of your session and don't receive an access code to resume it, please contact us. Your answers are probably not lost and there is hopefully no need to re-enter them from scratch.

**UPDATED LIST**

To give you some context for your answers, we are providing for each item of the updated list the corresponding item(s) in Round 1, the median score it obtained (with the interquartile range), as well as the percentage of participants scoring it 7 or above (i.e. recommending to include it). For full information about the scores, comments and actions taken in Round 1, please refer to the document "Round1\_per\_item\_results", attached here once again for your convenience.

As in Round 1, the research team is not providing any arguments to justify the items. Neutral clarifications are provided for some items, when the proposed concepts might not be common knowledge for all of the involved stakeholders or to clarify why certain edits have been made between the rounds. These clarifications are only informative and do not form part of the item.

[Attachment: "Round1\_per\_item\_results.pdf"]

---

**TITLE/ABSTRACT**

Title/abstract - Item 1a: Identify the study as early stage or formative clinical evaluation of an artificial intelligence or machine learning based decision support system, mentioning the clinical problem addressed.

Formerly item 1 (modified).

Median score = 8 (IQR: 6-9)

70.6% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---

Title/abstract - Item 1b: Provide a structured summary of the study, including: target clinical problem, intended use of the algorithm and integration in the clinical pathway, type of algorithm, study design, study setting, number of patients and users included, control group if applicable, primary and secondary outcomes, key safety endpoints, human factors aspects evaluated, main results, conclusions.

Formerly item 2 (modified).

Median score = 8 (IQR: 7-9)

79.1% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---



**INTRODUCTION**

Target clinical problem and population - Item 2: Describe the target clinical problem and medical condition, including the current state of the art practice, and the target patient population.

Formerly item 3 (modified).

Median score = 8 (IQR: 7-9)

88.3% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Intended use - Item 3: Describe the intended use of the algorithm, its planned integration in the care pathway and the impact in terms of patient outcomes it intends to achieve.

Formerly item 4 (modified).

Median score = 9 (IQR: 8-9)

93.3% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Current stage of development - Item 4: Describe the current stage of development of the algorithm (both from a scientific and a regulatory perspective). State if the algorithm is tested as a medical device and, if so, which regulatory approval is sought/was obtained.

Formerly item 7 (modified).

Median score = 7 (IQR: 6-8)

60.0% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Objectives - Item 5: State the study objectives.

Formerly item 8 (identical).

Median score = 9 (IQR: 8-9)

96.6% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

**METHODS**

Research governance - Item 6a: Provide a reference to any study protocol, study registration number and ethics approval.

Formerly item 9 (modified).

Median score = 8 (IQR: 7-9)

75.6% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Research governance - Item 6b: State what measures were taken to protect patient privacy and data security.

Formerly item 18 (identical).

Median score = 7 (IQR: 6-9)

64.2% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Study design - Item 7: Describe the study design.

Formerly item 11 (modified).

Median score = 8.5 (IQR: 7-9)

83.1% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Participants - Item 8a: Describe precisely how patients were recruited, stating the inclusion and exclusion criteria, and how the number of recruited patients was selected.

Formerly item 12...

Median score = 9 (IQR: 8-9)

92.5% of participants scoring  $\geq 7$

... and item 13 (merged and modified).

Median score = 8 (IQR: 6-9)

71.7% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---

Participants - Item 8b: Describe precisely how users were recruited, stating the inclusion and exclusion criteria, and how the number of recruited users was selected. If applicable, describe the control group in sufficient detail to allow replication.

Formerly item 12...

Median score = 9 (IQR: 8-9)

92.5% of participants scoring  $\geq 7$

... and item 13 (merged and modified).

Median score = 8 (IQR: 6-9)

71.7% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Participants - Item 8c: Describe any attempts to familiarise the users with the algorithm, including any training received.

Formerly item 24 (identical).

Median score = 8 (IQR: 7-9)

81.5% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Algorithm - Item 9: Briefly describe the algorithm, including: the version number, the type of AI model used, the characteristics of the patient population on which it was trained and the expected performance from in silico study. Refer to any previous development work.

Formerly item 5...

Median score = 8 (IQR: 7-9)

77.5% of participants scoring  $\geq 7$

... and item 6 (merged and modified).

Median score = 8 (IQR: 6-9)

71.7% of participants scoring  $\geq 7$

Comment from the research team: a detailed description of the algorithm is not expected, as this will be covered in previous development work. This item only requires a brief summary of its key characteristics so the algorithm can be appraised in the context of a clinical study.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Implementation - Item 10a: Describe precisely the environment in which the algorithm was tested, including the availability of the algorithm's input data and 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.

Formerly item 16 (modified).

Median score = 8 (IQR: 7-9)

84.9% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Implementation - Item 10b: Describe the clinical workflow/pathway in which the algorithm was deployed and who held the responsibility for the final clinical decision.

New item from the thematic analysis

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Implementation - Item 10c: Describe precisely how the algorithm was used and the timing of the decision support.

Formerly item 15 (modified).

Median score = 9 (IQR: 8-9)

91.5% of participants scoring  $\geq 7$

Comment from the research team: for example, decision support can be provided while users access clinical information and make their initial decision (first reader paradigm) or after users have made their initial decision (second reader paradigm).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Implementation - Item 10d: Describe the technical details of the implementation, including the integration within the existing study site IT infrastructure, the software and hardware needed to run the algorithm and any algorithmic thresholds used.

Formerly item 14 (split and modified).

Median score = 8 (IQR: 6-9)

71.7% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Implementation - Item 10e: 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.

Formerly item 17 (modified).

Median score = 8 (IQR: 7-9)

86.7% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Implementation - Item 10f: Describe the algorithm outputs and how they were presented to the users.

Formerly item 14 (split and modified).

Median score = 8 (IQR: 6-9)

71.7% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Outcomes - Item 11a: Specify the primary and secondary outcomes measured.

Formerly item 10 (identical).

Median score = 9 (IQR: 8-9)

87.5% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Outcomes - Item 11b: Describe how algorithm recommendation/output errors were defined and how they were identified.

Formerly item 23 (identical).

Median score = 8 (IQR: 7-9)

90.0% of participants scoring  $\geq 7$

Comment from the research team: For example, errors can be defined in comparison with a reference standard (not always available), as a failure to detect an event within a given timeframe, or as a therapeutic option vetoed by the clinical team.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Analysis - Item 12: Describe the pre-specified analysis plan for the primary and secondary outcomes as well as for any prespecified additional analyses, including subgroup analyses and their rationale.

Formerly item 20...

Median score = 8 (IQR: 7-9)

79.7% of participants scoring  $\geq 7$

... and item 21 (merged and modified).

Median score = 7 (IQR: 6-8)

70.1% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Safety - Item 13a: Define the algorithm safety requirements, how these were established preclinically, and how compliance to these requirements was evaluated during the study.

Formerly item 22 (identical).

Median score = 8 (IQR: 7-9)

84.0% of participants scoring  $\geq 7$

Comment from the research team: Safety requirements are generated by domain experts or derived from regulatory requirements and are informed by the system's risk analysis (AMLAS 2021). They should ensure an acceptable level of risk compared to the potential benefits of the system.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Safety - Item 13b: Describe the methodology used to detect any new, unexpected risks arising from the real-life clinical use of the algorithm.

New item from the thematic analysis.

Comment from the research team: while item 13a focuses on a theoretical/logical assessment of risk and compliance to already defined safety requirements, item 13b is about the empirical identification of any new risks emerging from the actual use of the algorithm in real life clinical settings, which were not or could not be identified during the preclinical risk assessment.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Human factors - Item 14: Describe the human factors tools, methods or frameworks used, the use cases considered and the users involved in the human factors evaluation.

Formerly item 25...

Median score = 7 (IQR: 6-8.5)

70.6% of participants scoring  $\geq 7$

... and item 26...

Median score = 8 (IQR: 6.5-9)

74.8% of participants scoring  $\geq 7$

... and item 38 (merged and modified).

Median score = 7 (IQR: 6-9)

66.9% of participants scoring  $\geq 7$

Comment from the research team: this is a general item about human factors methodology, supporting a more detailed set of items on specific human factors aspects to consider in the results section. The explanation paragraph for this item will include references to standard methods.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

Patient engagement - Item 15: State whether patients were involved in any aspect of the study design, conduct or in the development of the research question or outcome measures.

Formerly item 27...

Median score = 7 (IQR: 5.5-8)

54.6% of participants scoring  $\geq 7$

... and item 43 (merged and focus modified).

Median score = 7 (IQR: 5-8)

55.9% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

Ethics consideration - Item 16: Describe any ethics methodology, consultation or involvement during the design or implementation of the study.

New item from the proposed new item list.

Comment from the participant proposing this item: "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)

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

**RESULTS**

Participants - Item 17a: Describe the patient study group baseline characteristics (number, number of centres, age, sex, ethnicity if relevant, comorbidities, prevalence of the target conditions, etc.).

Formerly item 30 (identical).

Median score = 9 (IQR: 8-9)

95.8% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Participants - Item 17b: Describe the users study group baseline characteristics (number, number of centres, specialty, seniority, previous experience with digital support, etc.).

Formerly item 29 (identical).

Median score = 8 (IQR: 7-9)

87.5% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Implementation - Item 18a: Report on the user exposure to the algorithm (implementation reach), on the number of instances the algorithm was used (implementation dose) and on the users' adherence to the intended implementation (implementation fidelity).

Formerly item 31 (modified).

Median score = 8 (IQR: 7-9)

81.4% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

Implementation - Item 18b: Report changes caused by the algorithm to the clinical workflow, if any.

Formerly item 41 (modified).

Median score = 7 (IQR: 6-8)

70.3% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):



---

Modifications - Item 19: Report any changes made to the algorithm or its hardware platform between the prototype used at the beginning of the study and its final version. Report the timing of these modifications and the changes in outcomes observed after each of them.

Formerly item 42 (modified).

Median score = 8 (IQR: 7-9)

78.3% of participants scoring  $\geq 7$

See comment on page 3 of the executive summary.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Main results - Item 20a: Report on the prespecified outcomes for the algorithm-assisted users (both overall and at an individual user level), including any variation over time.

Formerly item 32 (split and modified).

Median score = 8 (IQR: 7-9)

89.0% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Main results - Item 20b: Report on the prespecified outcomes for the stand-alone algorithm, if applicable.

Formerly item 33 (modified).

Median score = 7 (IQR: 6-8)

66.4% of participants scoring  $\geq 7$

Comment from the research team: in other words, the hypothetical clinical performance of the "stand-alone" algorithm, without human intervention, in the study environment (if this data can be collected).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Main results - Item 20c: Report on the prespecified outcomes for the control group, if applicable.

Formerly item 32 (split and modified).

Median score = 8 (IQR: 7-9)

89.0% of participants scoring  $\geq 7$

See comment on page 3 of the executive summary.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Safety and errors - Item 21a: Report on the compliance with the specified safety requirements and any severe adverse events.

Formerly item 35 (identical).

Median score = 8 (IQR: 7-9)

82.1% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

Safety and errors - Item 21b: Report any additional risks identified from the real-life clinical use of the algorithm.

New item from the thematic analysis.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

Safety and errors - Item 21c: Report any algorithm malfunction or issues with hardware or software during the study.

New from the thematic analysis.

Comment from the research team: for example, the failure to load data from the patient record or to produce any output at all. This should be differentiated from algorithm recommendation/output errors.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

Safety and errors - Item 21d: Report any algorithm recommendation errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual impact on patient care.

Formerly item 34 (modified).

Median score = 9 (IQR: 7-9)

90.7% of participants scoring  $\geq 7$

See comment on page 3 of the executive summary.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

Safety and errors - Item 21e: Report any human errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual implication for patient care.

New from the thematic analysis.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Subgroup analysis - Item 22: Report on the difference in the main outcomes according to the specified subgroups.

New from the proposed new item list.

Comment from the research team: several defined subgroup analyses were proposed by participants (for example, based on the user's occupational experience or the socio-economic status of the patient) and would be mentioned as possible subgroups in the item's explanation.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Human factors - Item 23a: Report on the user agreement with the algorithm. Describe any instances of and reasons for user deviation from the algorithm's recommendations and, if applicable, user changing their mind based on the algorithm recommendations.

Formerly item 36 (modified).

Median score = 9 (IQR: 7-9)

88.3% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Human factors - Item 23b: Report on the evolution of users' trust in the algorithm.

Formerly item 37 (modified).

Median score = 7 (IQR: 6-8.25)

65.0% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Human factors - Item 23c: Report on the usability evaluation, including time to task completion, display interface evaluation and user satisfaction.

Formerly item 39 (modified).

Median score = 7 (IQR: 7-8)

75.6% of participants scoring  $\geq 7$

Comment from the research team: usability is defined as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 9241-11).

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Human factors - Item 23d: Report on the user workload and learning curves evaluation.

New item from the thematic analysis (workload), learning curves formerly in item 37 (median = 7, IQR:6-8.25, 65% of participant scoring 7 or above).

Comment from the research team: workload and learning curves have been merged into the same item because their interpretations are linked. For example, it is expected that the perceived workload will decrease as the users progress on their learning curve.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

---

Human factors - Item 23e: Report on the user perception of the algorithm outputs' interpretability and clinical value.

Formerly item 40 (modified).

Median score = 7 (IQR: 6-8)

64.7% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

---

Comment (optional):

---

**DISCUSSION**

Support intended purpose - Item 24: Discuss whether the obtained results support the intended purpose of the algorithm in real world clinical settings.

Formerly item 45 (modified).

Median score = 8.5 (IQR: 7-9)

86.7% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---

Safety and errors - Item 25: Discuss what the results suggest about the safety profile of the algorithm. Discuss the algorithm's errors and, if appropriate, identify any underlying pattern or algorithmic bias, explain how these can be mitigated.

Formerly item 47...

Median score = 9 (IQR: 7-9)

89.1% of participants scoring  $\geq 7$

... and item 48 (merged and modified).

Median score = 8 (IQR: 7-9)

90.0% of participants scoring  $\geq 7$

See comment on page 3 of the executive summary.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---

Human factors - Item 26: Discuss the results of the human factors evaluation and the reasons for human deviation from the algorithm's recommendations or intended use.

Formerly item 46...

Median score = 8 (IQR: 7-9)

85.8% of participants scoring  $\geq 7$

... and item 49 (merged and modified).

Median score = 7 (IQR: 6-8)

70.9% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---

---

Scale up - Item 27: Discuss the scale-up feasibility and requirements, as well as the possible design of large-scale summative evaluation in light of the obtained results. Summarise the lessons learned from the study.

Formerly item 50...

Median score = 7 (IQR: 6-8)

59.7% of participants scoring  $\geq 7$

... and item 51...

Median score = 7.5 (IQR: 7-8.25)

78.3% of participants scoring  $\geq 7$

... and item 52 (merged and modified).

Median score = 8 (IQR: 7-9)

82.4% of participants scoring  $\geq 7$

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5   ☐ 6   ☐ 7   ☐ 8   ☐ 9   ☐ I don't know

---

Comment (optional):

---

Strengths and limitations - Item 28: Discuss the strengths and limitations of the study, including any bias in the study design.

New item from the proposed new item list.

Comment from the participant proposing the item: "This is a standard item for the Discussion of any empirical study but it is nevertheless usually included in reported standards."

☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5   ☐ 6   ☐ 7   ☐ 8   ☐ 9   ☐ I don't know

---

Comment (optional):

**STATEMENTS**

Conflicts of interest - Item 29: Disclose any relevant conflict of interest, including: the source of funding for the study, the role of funders, any other role played by commercial companies and authors' conflicts of interest.

Formerly item 53 (modified)

Median score = 9 (IQR: 8-9)

97.5% of participants scoring  $\geq 7$

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---

Data availability - Item 30: Disclose if and how data and code (pre-processing and algorithm) are available.

Formerly item 54 (modified).

Median score = 8 (IQR: 7-9)

80.3% of participants scoring  $\geq 7$

Comment from the research team: this item does not aim to enforce the disclosure of the code or underlying data. Instead, it is about reporting if, and under which conditions, the code and data could be accessed by other research teams to reproduce the results.

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7 ☐ 8 ☐ 9 ☐ I don't know

Comment (optional):

---

**ADDITIONAL QUESTIONS**

1. When compared with the original item list in Round 1, do you think the second iteration of the list (updated list in Round 2) was qualitatively better?

- ☐ Yes  
☐ No  
☐ Similar  
☐ I don't know

The two lists are attached here below for your convenience if needed.

Original list

[Attachment: "Round1\_original\_item\_list.pdf"]

Updated list

[Attachment: "Round2\_updated\_item\_list.pdf"]

2. Based on your answers to Round 2, how many items/sub-items do you think the final list should comprise (where for example 1a and 1b would be 2 items/subitems for the purposes of a total count)?

\_\_\_\_\_

3. Do you have any additional comments on Round 2?

\_\_\_\_\_

4. As mentioned in the participant information sheet, the names and affiliations of the Delphi participants will be disclosed in the final publication, both for transparency and to acknowledge your contribution. Could you please provide the name(s) of the institution(s) you would like to appear under in the participant list (ideally in the format: institution, city, country)?

\_\_\_\_\_

PERSONAL INFORMATION (additional questions for participants who did not complete Round 1)

5. In which country do you mainly work?

\_\_\_\_\_



6. Which of the following groups of stakeholders would describe you best? (multiple answers possible)

- ☐ Administrator/other management position in hospital
- ☐ Allied health professional
- ☐ Clinician
- ☐ Engineer/Computer scientist
- ☐ Entrepreneur
- ☐ Epidemiologist
- ☐ Ethicist
- ☐ Funder
- ☐ Human factors specialist
- ☐ Implementation scientist
- ☐ Journal editor
- ☐ Methodologist
- ☐ Patients' representative
- ☐ Payer/Commissioner
- ☐ Policy maker/official institutions representative
- ☐ Private sector representative
- ☐ Psychologist
- ☐ Regulator
- ☐ Statistician
- ☐ Trialist
- ☐ other

If other, please specify.

---

You mentioned (either during this or the previous round) that you have a private sector affiliation or are developing your own company/companies as an entrepreneur. For transparency, could you please name the commercial entities you are linked with (even if already stated as main affiliations) and the nature of this relationship (for example, MyCompany, CMO and equity holder)?

---

You mentioned (either during this or the previous round) that you have a private sector affiliation or are developing your own company/companies as an entrepreneur. For transparency, could you please name the commercial entities you are linked with (even if already stated as main affiliations) and the nature of this relationship (for example, MyCompany, CMO and equity holder)?

---

7. Could you please briefly describe your level of experience/type of expertise with Artificial Intelligence or Machine Learning?

---

8. Could you please briefly describe your level of experience/type of expertise with clinical evaluation or technology implementation?

---