

DECIDE-AI

Round 1 analysis – executive summary

29.04.21

Data collection

- 187 experts invited, of which 117 accepted the invitation (acceptation rate = 63%) + 21 experts contacted the research team = 138 Delphi Round 1 questionnaires sent
- 121 completed questionnaires received (response rate = 88%), with 121 sets of free text answers and 120 sets of items scores included in the analysis (the item scores from 1 participant were excluded due to suspicion of scale inversion)
- 46033 words of narrative answers to the first four open-ended questions and general comments question, 6480 item scores, 312 item or section specific comments, 64 proposed new items.

Analysis and results

The following analyses were performed:

- Thematic analysis of all free text answers to the open-ended questions
- Basic statistical analysis (median, IQR, mean, SD, % scoring ≥ 7 , % scoring ≤ 3) of all item scores, including stakeholder subgroup analysis
- Comments grouped by item and comment summary produced for each item
- Case by case analysis of each proposed new item

All analyses were performed in parallel by two members of the research team (Myura Nagendran and Baptiste Vasey). Conflicts were resolved in consensus. The original item list was updated taking into consideration the results of the analyses mentioned. 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. In the updated item list:

- 9 items were kept without modification
- 22 items were reworded/completed
- 22 items were reorganised (merged/split), becoming 13 items
- 2 items were dropped
- 9 new items were added (3 from the proposed new item list, 6 from the thematic analysis based on the answers to the open-ended questions)
- 22 new item components were included in existing items (9 from the proposed new item list, 10 from the thematic analysis, 3 from both)

Interpretation and discussion

The thematic analysis yielded two main results. First, it showed that the original item list proposed by the Steering Committee was in keeping with the overall expectations of the expert cohort. Indeed, almost all the themes mentioned in the original item list were also found numerous times in the answers to the four open-ended questions. When interpreting the thematic analysis, it should be kept in mind that these themes were highlighted by participants before they were exposed to the original item list. It could be argued that some of the questions were directed (e.g. one question was directly asking about human factors, hence it was unsurprising that many participants mentioned human factors aspects in their answers). However, many themes appeared in the answers to several of the questions, supporting the proposition that they were not induced entirely by a directed formulation of the question (e.g. human factors aspects were also mentioned in the answers to the first question, before the participants were prompted to think about human factors). Secondly, it provided 6 new items which had not been identified in the original item list.

Most items were ranked 7 and above in importance and our strategy for Round 1 did not allow a reduction in the number of items. The number lost due to items being merged or dropped was compensated for by new items proposed by the participants or identified from the thematic analysis. The updated list is therefore still too long (53 items/subitems) for a final reporting guideline.

During the Round 1 analysis of comments for each item, the research team took into account that many participants (usually over 100) submitted scores but no comments. This could reflect approval, or a lack of interest for a specific item. We tried to keep a balance between comments (mostly critical) and score consensus (mostly supportive) when drafting the updated item list. This is the reason why not all comments were acted upon. A substantial number of general positive comments about the current item list were also found during the analysis which have not been reported in the comment summary for brevity.

The two items about health economic assessment were dropped following the recommendation of several participants and considering the low consensus obtained in Round 1.

Reflecting on all the answers received in Round 1, the most commonly cited areas of uncertainty or disagreement were:

- the main objective of early-stage clinical evaluation (should effectiveness be considered and if so, how?)
- the timeline of usability and safety testing (preclinical or early clinical)
- the role played by control groups (is a control group needed in early formative study?)
- the notion of changes applied to the algorithm (need for rapid iterative design vs. stability of the intervention)
- the reporting and analysis of errors (case-by-case description vs. pattern/bias analysis).

When updating the item list and based on participant feedback, the research team adopted the following positions, which are open to comment under the relevant items in Round 2.

- a) The early-stage clinical evaluation is a stepping-stone toward definitive summative evaluation. As such, it aims to analyse the clinical performance of the algorithm and assisted humans in real world settings, consolidate the safety and usability testing already initiated during preclinical development and collect the information necessary for the design of larger-scale trials. Given the (likely) small study group and study design often adopted for this type of early study, conclusions about effectiveness cannot be drawn at this stage, although some preliminary information (expected effect size, subgroup difference, etc.) can be obtained to inform future study design. Therefore, control groups might be used (but are not indispensable) and if used, should be reported in sufficient detail.
- b) Since most early-stage evaluations will not be designed to make definitive claims about effectiveness, rapid prototyping (design changes or software update) is seen by many as important components of this type of study, although they may introduce additional complexity in the interpretation of results. Whether one agrees or not with the benefit of rapid prototyping, information about whether and how it happened is important to appraise the study results. Therefore, the item about reporting of changes was kept in the updated list.
- c) Detailed case-by-case analysis of errors will only be practical with a small number of occurrences, while the analysis of patterns of errors/algorithmic bias will only be meaningful with sufficient numbers of cases. Therefore, the items about errors were rephrased to shift the detailed description from individual occurrences to individual reasons, responses and impacts on patients. “If appropriate” was added to the item about pattern of errors and algorithmic bias to account for variable study sizes.

The first round of Delphi produced a rich database of opinions about early-stage clinical evaluation of AI-based decision support systems. We tried to analyse and integrate these opinions into the original item list and have now created an updated item list, which will be presented to the Delphi participants in Round 2. This second round will be important to assess the quality of the changes made by the research team, produce an overview of the current level of consensus and provide indications to the Consensus Group to assist them in the selection and wording of the final items for the DECIDE-AI reporting guidelines.