

DECIDE-AI

Round 1 per item results

26.04.21

TABLE OF CONTENTS

HOW TO READ THIS DOCUMENT	3
TITLE.....	4
ITEM 1.....	4
ABSTRACT.....	5
ITEM 2.....	5
INTRODUCTION	6
ITEM 3.....	6
ITEM 4.....	7
ITEM 5.....	8
ITEM 6.....	9
ITEM 7.....	10
ITEM 8.....	11
METHODS.....	12
ITEM 9.....	12
ITEM 10.....	13
ITEM 11.....	14
ITEM 12.....	15
ITEM 13.....	16
ITEM 14.....	17
ITEM 15.....	18
ITEM 16.....	19
ITEM 17.....	20
ITEM 18.....	21
ITEM 19.....	22
ITEM 20.....	23
ITEM 21.....	24
ITEM 22.....	25
ITEM 23.....	26
ITEM 24.....	27
ITEM 25.....	28
ITEM 26.....	29
ITEM 27.....	30
ITEM 28.....	31

RESULTS	32
ITEM 29.....	32
ITEM 30.....	33
ITEM 31.....	34
ITEM 32.....	35
ITEM 33.....	36
ITEM 34.....	37
ITEM 35.....	38
ITEM 36.....	39
ITEM 37.....	40
ITEM 38.....	41
ITEM 39.....	42
ITEM 40.....	43
ITEM 41.....	44
ITEM 42.....	45
ITEM 43.....	46
ITEM 44.....	47
DISCUSSION.....	48
ITEM 45.....	48
ITEM 46.....	49
ITEM 47.....	50
ITEM 48.....	51
ITEM 49.....	52
ITEM 50.....	53
ITEM 51.....	54
ITEM 52.....	55
STATEMENTS.....	56
ITEM 53.....	56
ITEM 54.....	57
SECTION SPECIFIC COMMENTS	58
INTRODUCTION.....	58
METHOD	58
RESULTS	58
DISCUSSION	58

How to read this document

This document was produced using data collected from the submitted scores and item/section specific comments. For each item, the following pieces of information are provided:

1. The number of comments made for the item. This helps to weight the opinions presented against the overall number of participants.
2. A summary of the comments submitted. In the context of this first round of Delphi, it was decided to report a summary (rather than every individual comment) in order to provide a balanced view of the different opinions. Please note that the comments are those of the participants and do not necessarily represent the opinion of the research team.
3. A visual representation of the submitted scores, broken down by stakeholder group, which should allow for a more granular interpretation of the current level of consensus. Please note that a same participant can belong to several stakeholder groups (according to their own answer to the related question in Round 1).
4. Scores summary statistics for the overall Delphi expert group, including a pointer to the stakeholder groups whose median score differs of two points or more from the overall median.
5. The action taken between Round 1 and 2. This include:
 - a. no change to the item
 - b. a change of wording
 - c. a merge or split (or both) with another item
 - d. the transfer of a specific concept to another item
 - e. discarding the item.
6. The number of the corresponding item(s) in the updated list. This number can significantly differ as the research team reorganised the item order after incorporating the feedback from Round 1.

Section specific comments are comments which could not be attributed to an individual item.

TITLE

Item 1

Identify the study as early-stage, exploratory or first-with-human clinical evaluation of an artificial intelligence or machine learning based decision support algorithm.

Number of comments: 32

Summarised participant comments:

- Ideally some standardized, broadly accepted wording
- Should include clinical problem, intended goal and intended use
- Should include study design and methodology
- Should mention the “stepping stone” stage of development
- “First-with-human” maybe too prescriptive (are silent/shadow trials first in human? What about studies testing new applications of existing algorithm?)
- “First-with-human” maybe too vague to non-medics, “exploratory” is too vague
- Study qualifiers should not be used in title
- Consider clarifying if referring to the algorithm or the intervention (could be an established algorithm used in a new context/for a new indication)
- Clarify in the abstract if title can’t include everything.

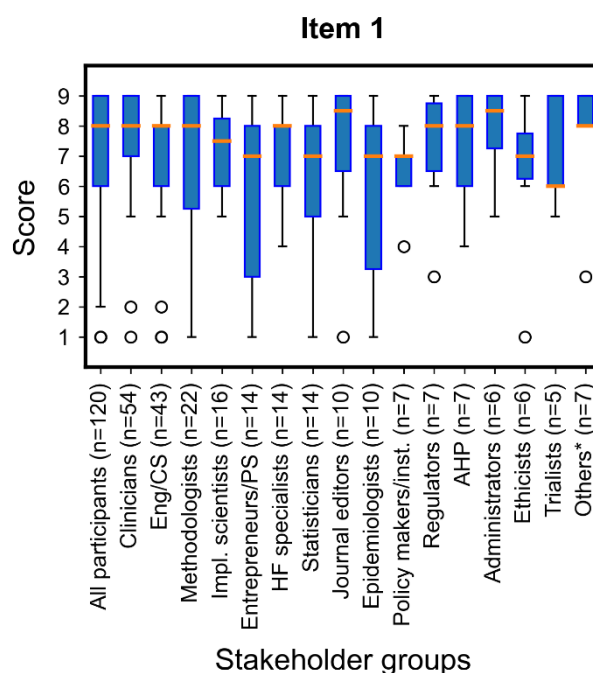


Figure 1: median score and IQR for item 1 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Alternative denominations proposed:

“Augmented intelligence study”, “Human-with-AI study”, “Human-with-AI Interaction, Refinement and Performance”, “Phase beta study of...”, “human-in-the-loop assessment”, “usability study”, “First with-human assessment of a machine learning-based clinical decision support system”.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% “I don't know”	Stakeholder group with median <=2 or >=2 points away from overall median
8	6	9	7.2	2.0	70.6	6.7	0.8	Trialists

Action taken: wording modified, becomes item 1a in the updated list.

ABSTRACT

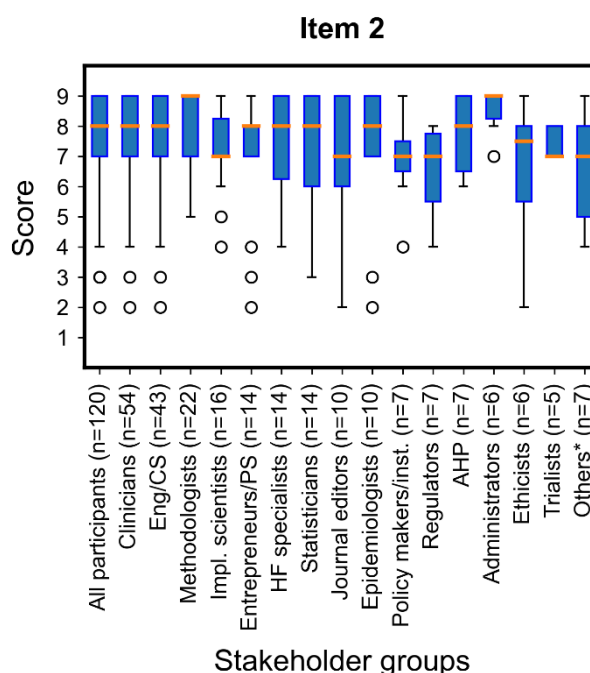
Item 2

Provide a structured summary of the study, including mention of: [will be completed according to the outcomes of the Delphi]

Number of comments: 25

Summarised participant comments:

- Need to re-ask in round 2 as the item is not yet complete
- Can go along the lines of normal abstracts but with a bit more attention on the data/population description
- State that it's a small-scale assessment prior to larger scale implementation
- Important but does not need a specific structure
- Specific guidance for abstract will need to be developed anyway
- Could use the same item as for CONSORT/SPIRIT-AI.



Proposed subitem to be included:

clinical problem addressed, intended use, model description, on-line vs. off-line, data set sizes (train, validate, test), cross-validation, input data, prespecified thresholds, ethical approval, status of algorithm used (in-house/commercial), healthcare and study settings, study design, study population, number of participants, control group, results on primary outcomes, key safety issues and endpoints, key user interaction issues, key outcomes, performance metrics, results from decision-curve analysis, summary of discrepancy between reported in-silico performance (particularly with respect to patient subpopulations), human factors approach, usability assessment, clinician trust, occurrence of disagreement and deviation from the recommendations, learning curves, conclusions.

*Figure 2: median score and IQR for item 2 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factor, AHP = Allied health professionals, *Patient representatives and psychologists.*

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	7.6	1.6	79.1	2.6	4.3	None

Action taken: missing part completed, becomes item 1b in the updated list.

INTRODUCTION

Item 3

Describe the target conditions and the patient population that would benefit from the algorithm, including information on the target conditions' prevalence and their impact on the healthcare system.

Number of comments: 3

Summarised participant comments:

- Clinical problem need to be clearly stated
- Could merge with item 4.

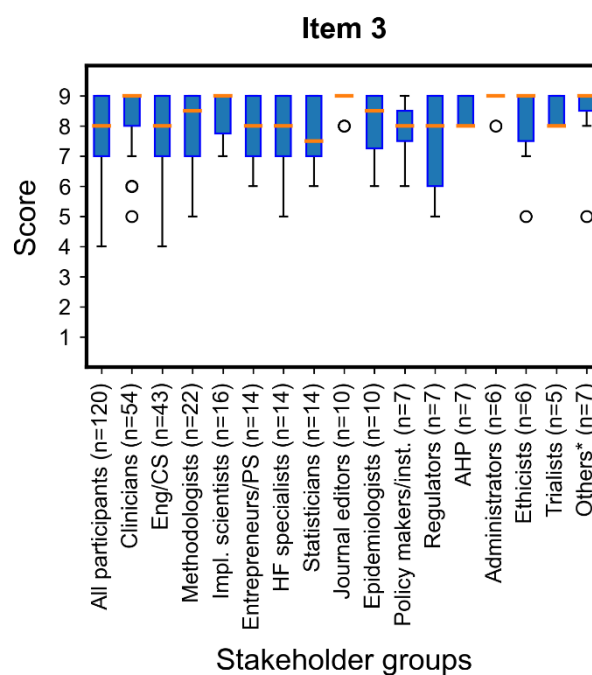


Figure 3: median score and IQR for item 3 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	8	1.2	88.3	0	0	None

Action taken: wording modified, becomes item 2 in the updated list.

Item 4

Describe the intended use of the algorithm, including its position in the care pathway and the conditions under which it would be used, the impact in terms of patient care it intends to achieve and the current state of the art practice.

Number of comments: 6

Summarised participant comments:

- Intended use essential, others not
- Is this right place for intended use?
- A description of the actions expected to be taken based on the algorithm outcomes should be included
- Anticipated place of use should be included
- Consider splitting the points about position and impact
- Explicitly describe the planned AI-human interaction
- Could merge with item 3.

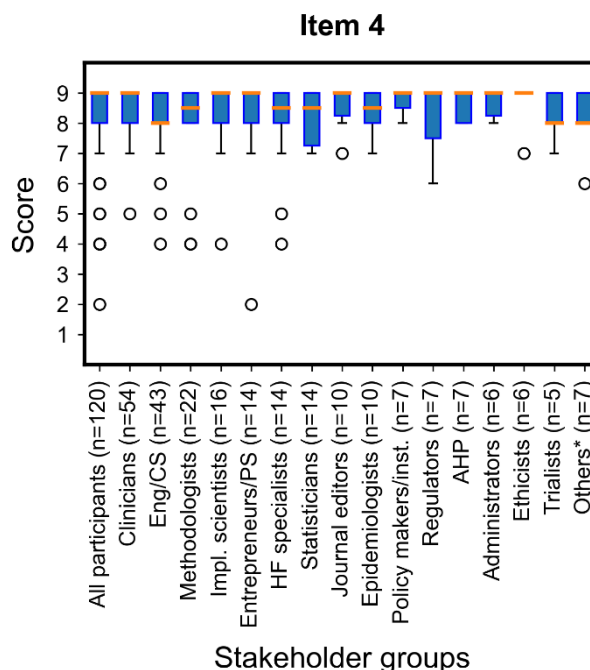


Figure 4: median score and IQR for item 4 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
9	8	9	8.2	1.3	93.3	0.8	0	None

Action taken: wording modified, becomes item 3 in the updated list, point about state of the art practice moved to item 2 in the updated list.

Item 5

Refer to the algorithm's development and validation studies and name the algorithm, including the version number. State the algorithm's expected performance from development and validation studies.

Number of comments: 12

Summarised participant comments:

- Could merge with item 6
- Move to methods
- Type of AI should be mentioned
- Cite previous work if published rather than repeating
- the expected algorithm performance should be mentioned
- Prior performance data may not be relevant (could create a false narrative, development/training data etc. rarely influences how "valid" the study is)
- Regulatory status should be mentioned.

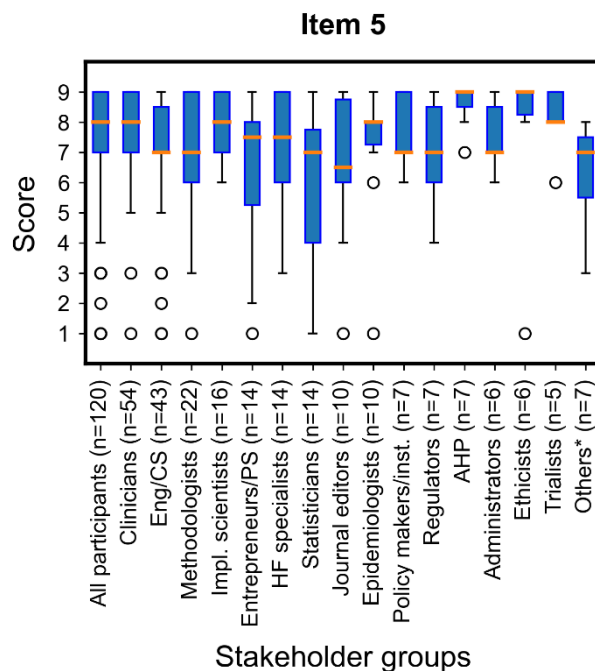


Figure 5: median score and IQR for item 5 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.3	1.9	77.5	5.8	0	None

Action taken: wording modified, moved to methods section, merged with item 6 of the original list, becomes item 9 in the updated list.

Item 6

Identify the dataset used to develop the algorithm and provide information on its relevance to the test environment, including the target conditions' prevalence when appropriate.

Number of comments: 5

Summarised participant comments:

- Could merge with item 5
- Move to methods
- Some training data may be protected but should nonetheless be discussed
- Cite previous work if published rather than repeating
- Differences/overlaps between the training set population and clinical evaluation population should be highlighted.

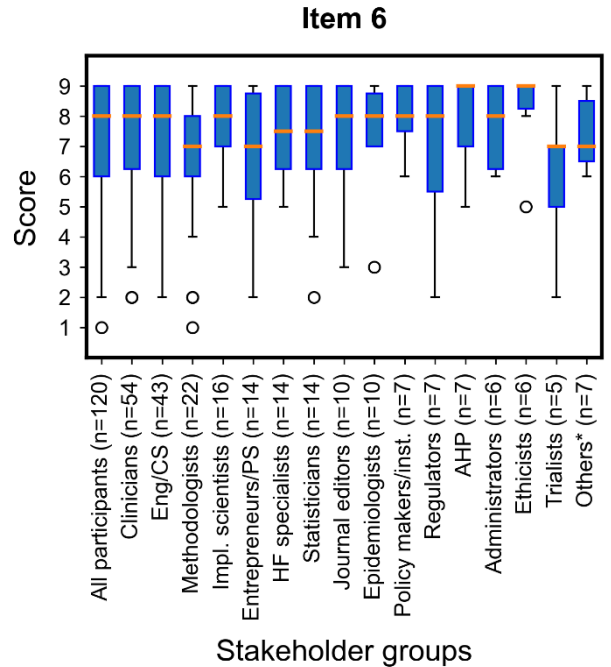


Figure 6: median score and IQR for item 6 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	6	9	7.3	1.9	71.7	5.8	0	None

Action taken: wording modified, moved to methods section, merged with item 5 of the original list, becomes item 9 in the updated list.

Item 7

Describe the current stage of development of the algorithm in terms of the essential questions which have been and remain to be answered about it (both from a scientific and a regulatory perspective).

Number of comments: 7

Summarised participant comments:

- Should include a clear statement whether the tested algorithm is a medical device or not (this could have an impact on additional requirements for the study)
- Regulatory environment is fast moving and may be out of date by the time the manuscript is published
- Not the right place for discussing this
- too vague
- cite previous work if published rather than repeating.

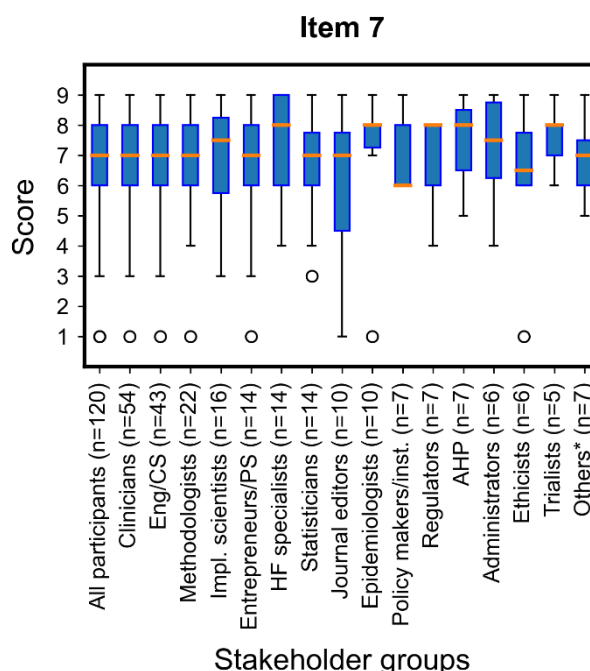


Figure 7: median score and IQR for item 7 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	6	8	6.9	1.8	60	5	0	None

Action taken: wording modified, becomes item 4 in the updated list.

Item 8

State the study objectives.

Number of comments: 2

Summarised participant comments:

- Could be more specific (e.g. provide suggestions)
- The study is either hypothesis testing or not. If not it does not contribute to determine impact on patient outcomes.

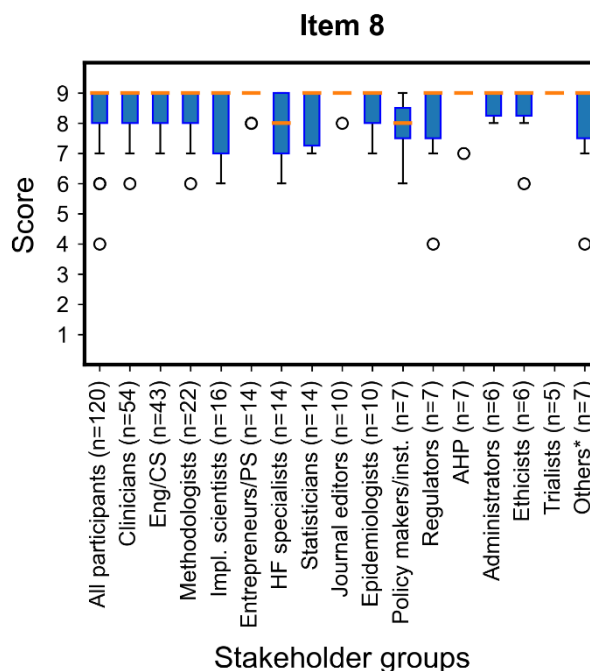


Figure 8: median score and IQR for item 8 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
9	8	9	8.5	0.9	96.6	0	0.8	None

Action taken: becomes item 5 in the updated list.

METHODS

Item 9

Provide a reference to any study protocol.

Number of comments: 5

Summarised participant comments:

- Maybe premature to expect protocol since the goal of this step of study is to refine the intervention itself
- Important to encourage pre-specified protocol
- State if the study was reviewed by a research ethics review committee
- Registration of the study should be encouraged (and could increase adherence to the guideline as it is unlikely that studies will have included all relevant reporting items unless already used during study design).

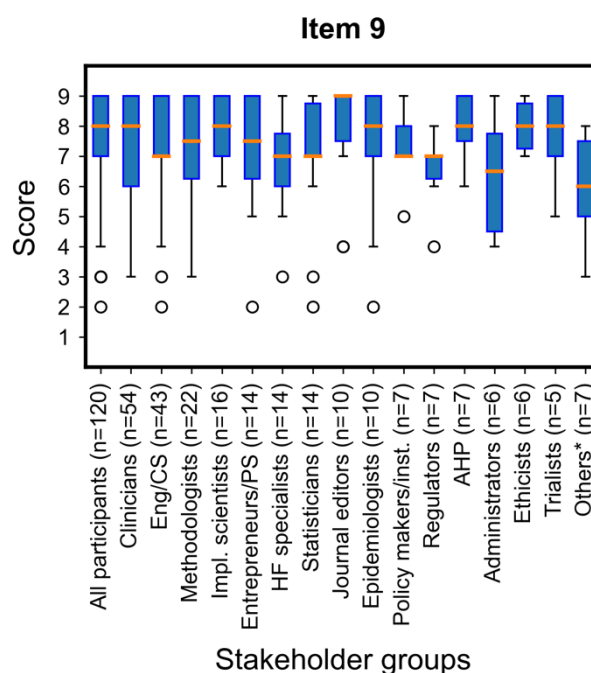


Figure 9: median score and IQR for item 9 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=8 points away from overall median
8	7	9	7.3	1.6	75.6	3.4	0.8	Others

Action taken: wording modified, becomes item 6a in the updated list.

Item 10

Specify the primary and secondary outcome measures.

Number of comments: 1

Summarised participant comments:

- Consider specifying that outcome measures are clinical.

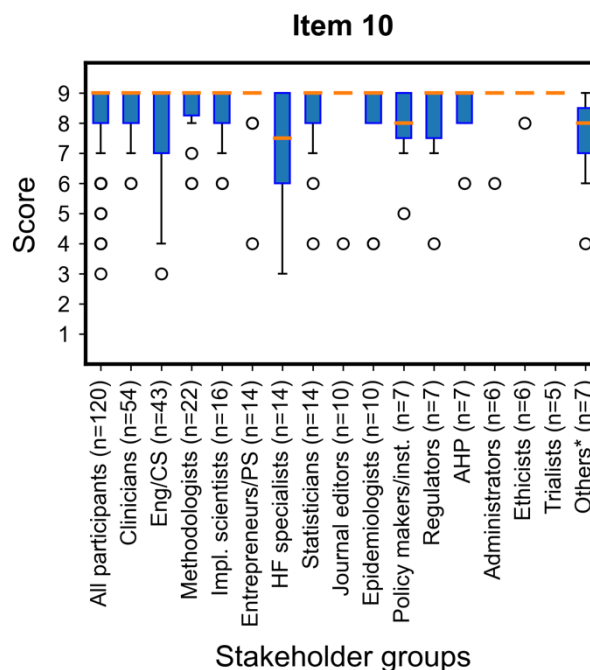


Figure 10: median score and IQR for item 10 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
9	8	9	8.2	1.3	87.5	0.8	0.0	None

Action taken: becomes item 11a in the updated list.

Item 11

Describe the study design using standard methodological terminology.

Number of comments: 2

Summarised participant comments:

- Vague
- Not clear if there is agreed standard methodological terminology.

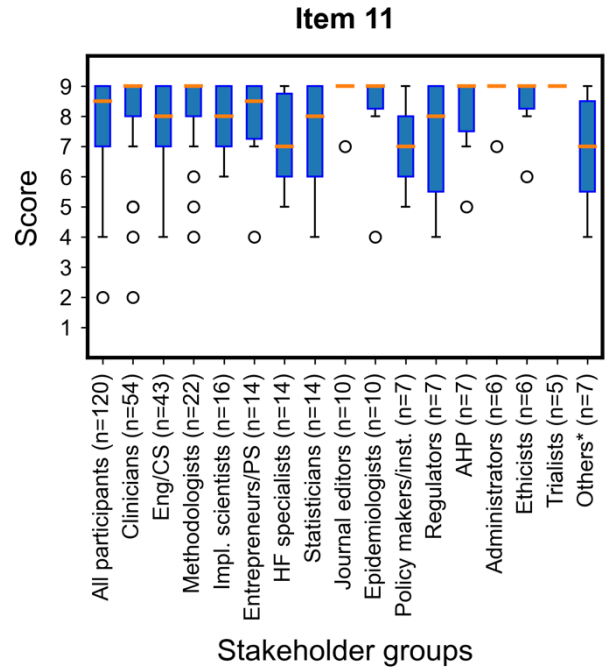


Figure 11: median score and IQR for item 11 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8.5	7	9	7.8	1.5	83.1	0.8	1.7	None

Action taken: wording modified, becomes item 7 in the updated list.

Item 12

Describe precisely how users and patients were selected. If only a subgroup of users took part in the human factors evaluation, describe how these were selected.

Number of comments: 3

Summarised participant comments:

- Change selected to recruited
- The part on subgroup of users taking part in the human factors evaluation is unclear
- Unit of assignment should be specified (the unit being assigned to use the algorithm).

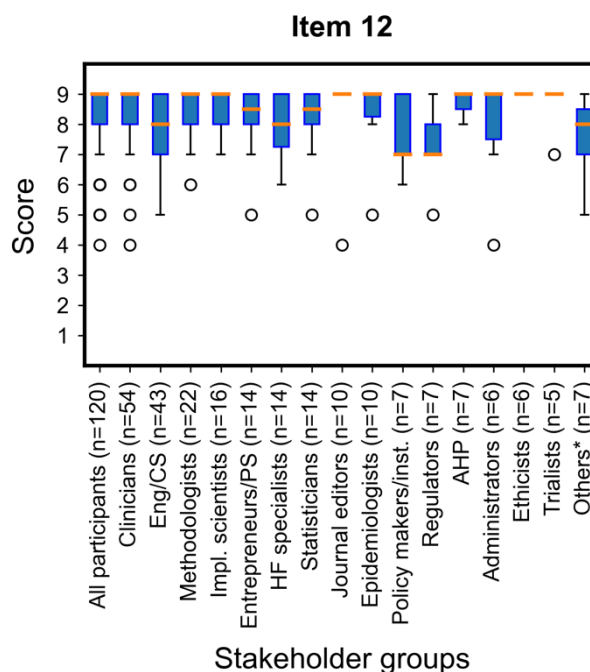


Figure 12: median score and IQR for item 12 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
9	8	9	8.2	1.1	92.5	0.0	0.0	Policy makers, Regulators

Action taken: wording modified, merged with item 13 of the original list, split into item 8a and 8b in the updated list, point about human factors evaluation subgroup moved to item 14 in the updated list.

Item 13

Justify the sample sizes (for both users and patients).

Number of comments: 3

Summarised participant comments:

- Not relevant for all types of study at this stage (Is this relevant if usability study? Will it be known in advance?) More relevant for final RCT.
- Justify might be replace by explanation about the number of included participants
- Sample size calculation are also possible for small-scale pilot studies, see <https://pubmed.ncbi.nlm.nih.gov/26146089>.

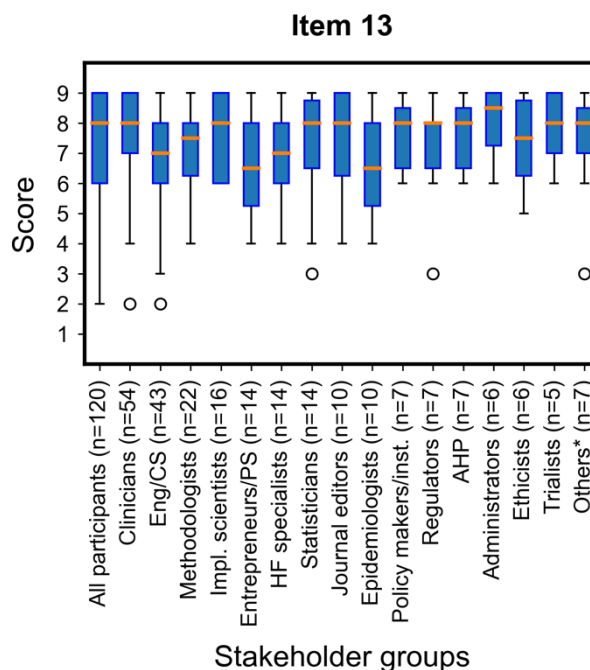


Figure 13: median score and IQR for item 13 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	6	9	7.3	1.7	71.7	2.5	0.0	None

Action taken: wording modified, merged with item 12 and 19 of the original list, split into item 8a and 8b in the updated list.

Item 14

Identify the hardware and software platforms used during the study. Describe the data needed by the algorithm as inputs, the data provided by the algorithm as outputs and the minimal computational resources needed.

Number of comments: 7

Summarised participant comments:

- Computing requirements may be of varying importance in different applications so use discretion
- Overlap with 16 and 17 regarding data, could be merged
- Item should be split
- Describe the deployment of the algorithm within the existing clinical infrastructure.

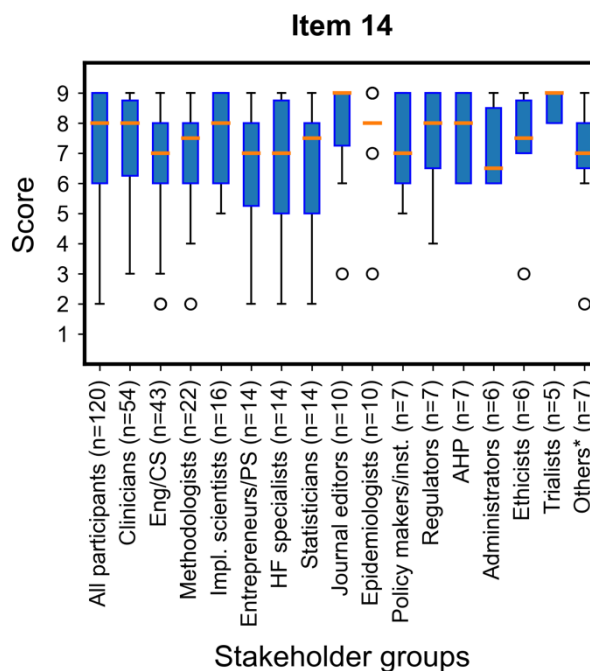


Figure 14: median score and IQR for item 14 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	6	9	7.2	1.8	71.7	5.0	0.0	None

Action taken: wording modified, split into items 10d and 10f in the updated list, point about data needed as inputs moved to item 10e in the updated list.

Item 15

Describe how the algorithm was used, at which stage of the decision-making process and who held the responsibility for the final clinical decision.

Number of comments: 4

Summarised participant comments:

- Could be merged with 16
- shared decision making with patients important to consider (rather than paternalistic clinician only, could influence outcomes independently of the algorithm)
- Describe how the information were presented, including explanation given
- Need to be detailed enough to allow replication.

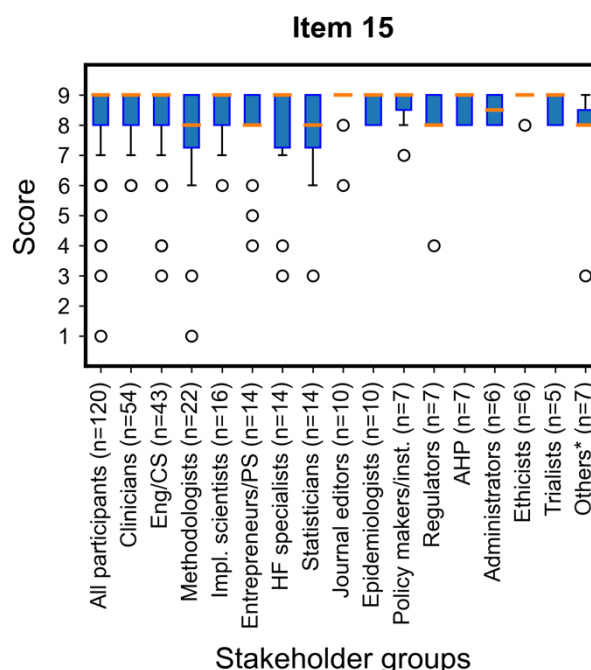


Figure 15: median score and IQR for item 15 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
9	8	9	8.2	1.3	91.5	1.7	1.7	None

Action taken: wording modified, becomes item 10c in the updated list, point about responsibility moved to item 10b.

Item 16

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.

Number of comments: 6

Summarised participant comments:

- Could merge with 15
- Overlap with 14 and 17 regarding data
- Specify what is meant by input data
- Input data could be protected/unavailable but can still discuss it
- Availability of algorithm input data should be in results
- Rephrase second part to *“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”*.

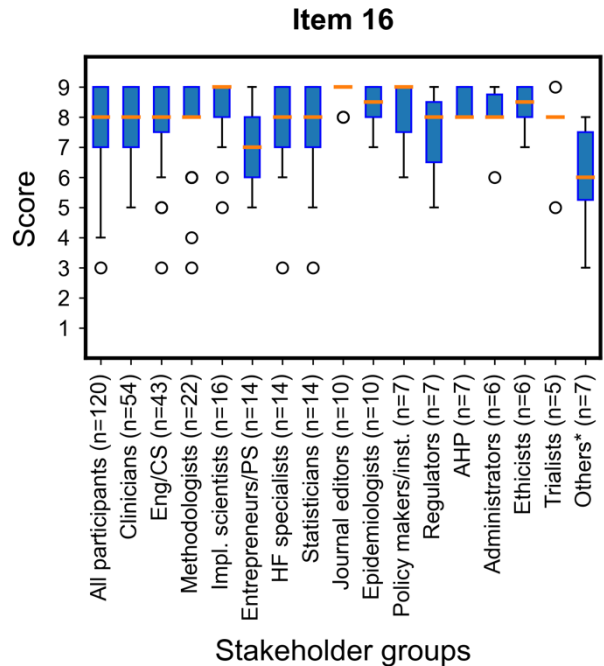


Figure 16: median score and IQR for item 16 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	7.9	1.3	84.9	0.8	0.8	Others

Action taken: wording modified according to comment, becomes item 10a in the updated list.

Item 17

Describe how the patient data were acquired (including from which sources), how they were processed and how missing or low-quality data were handled.

Number of comments: 5

Summarised participant comments:

- Overlap with 14 and 16 regarding data
- Too early in the evaluation to evaluate this
- Description of data pre-processing and validation is important
- Patient data is vague -> does it refer to follow-up?
- Describe precisely the equipment used to acquire the data.

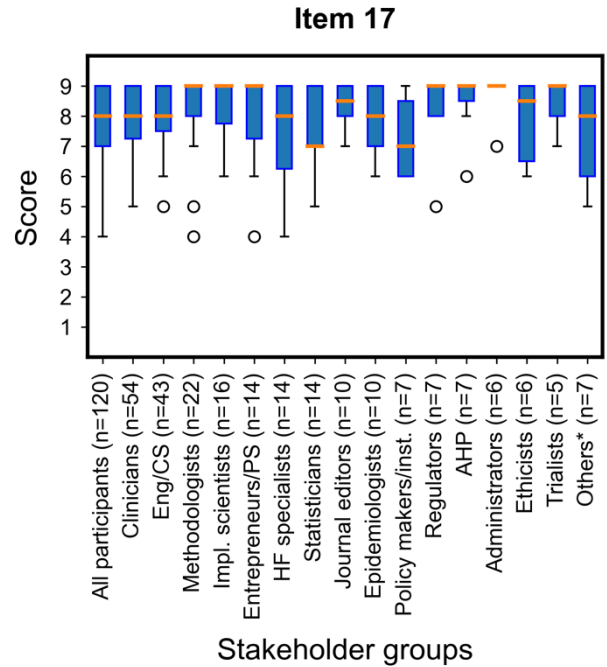


Figure 17: median score and IQR for item 17 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	7.9	1.2	86.7	0.0	0.0	None

Action taken: wording modified, becomes item 10e in the updated list.

Item 18

State what measures were taken to protect patient privacy and data security.

Number of comments: 3

Summarised participant comments:

- Generic issue and usually covered by ethics, instead better to just state ethics approval obtained as an item.

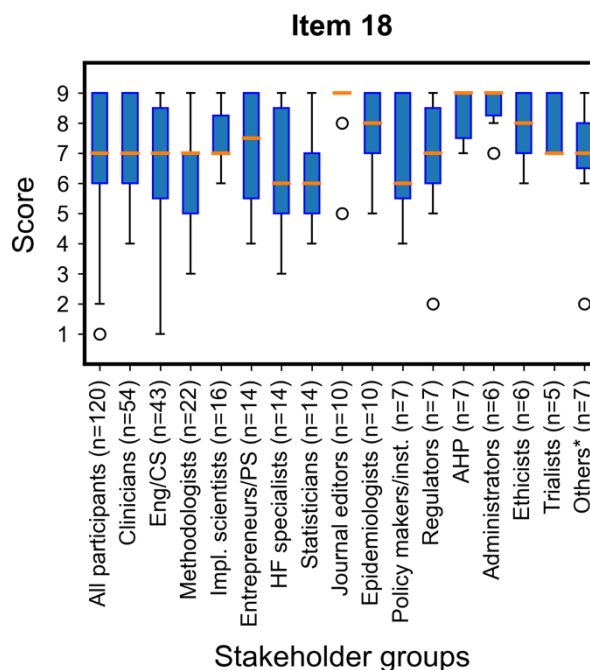


Figure 18: median score and IQR for item 18 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	6	9	7.0	1.8	64.2	3.3	0.0	Journal editors, Allied health professionals, Administrators

Action taken: becomes item 6b of the updated list.

Item 19

Describe the control group in sufficient detail to allow replication.

Number of comments: 8

Summarised participant comments:

- There may not be a control group for all studies at this stage of the evaluation, add '*if applicable*'
- Unclear exactly what is meant by control group.

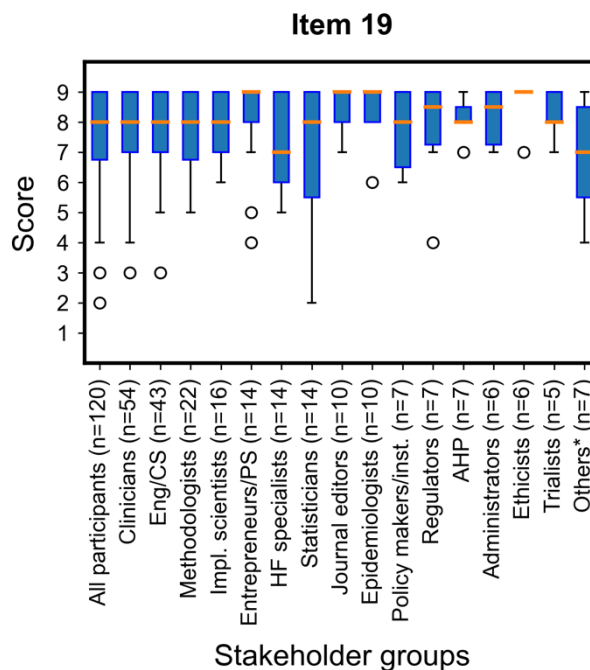


Figure 19: median score and IQR for item 19 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	6.75	9	7.5	1.6	75.0	1.7	3.4	None

Action taken: added *if applicable*, merged with items 13 of the original list, becomes item 8b in the updated list.

Item 20

State the predefined statistical analysis plan and any additional exploratory analyses performed (state if the chosen approach accounts for both user and patient variability).

Number of comments: 4

Summarised participant comments:

- Merge with 21
- Important to mention variation across users
- There is no legal requirement for predefined analysis plan in MDR/IVDR, depends on the use case.

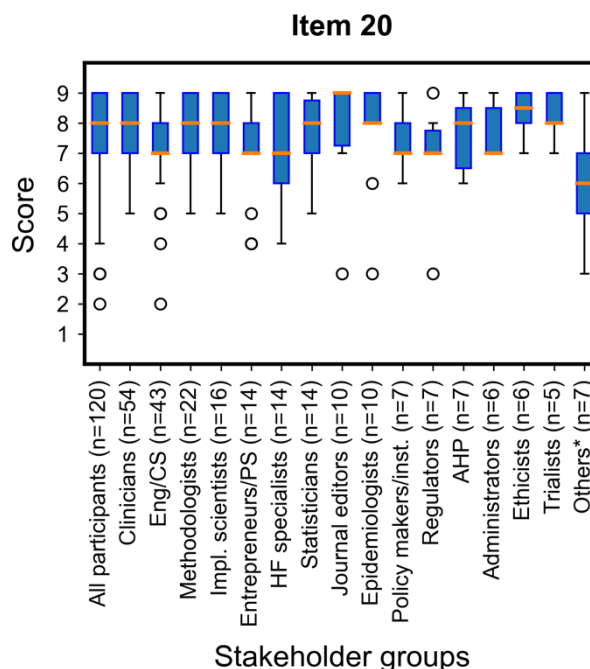


Figure 20: median score and IQR for item 20 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	7.4	1.5	79.7	2.5	1.7	Others

Action taken: wording modified, merged with item 21 of the original list, becomes item 12 in the updated list.

Item 21

State any pre-specified subgroup analyses and their rationale.

Number of comments: 6

Summarised participant comments:

- Merge with 20
- More detail on a priori identification of subgroup of interest
- Should be stronger in term of identifying vulnerable subgroups and protected attributes as well as testing for unfairness between subgroups
- Safeguarding against p-fishing with pre-defined subgroup analysis is less of a concern at the early trial stage.

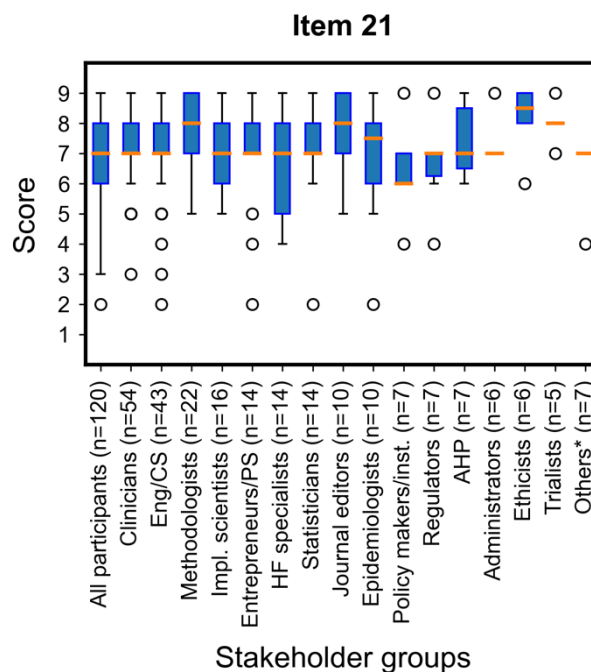


Figure 21: median score and IQR for item 21 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	6	8	7.1	1.5	70.1	2.6	2.6	None

Action taken: wording modified, merged with item 20 of the original list, becomes item 12 in the updated list.

Item 22

Define the algorithm safety requirements, how these were established and how compliance to these requirements was evaluated.

Number of comments: 0

Summarised participant comments:

- None

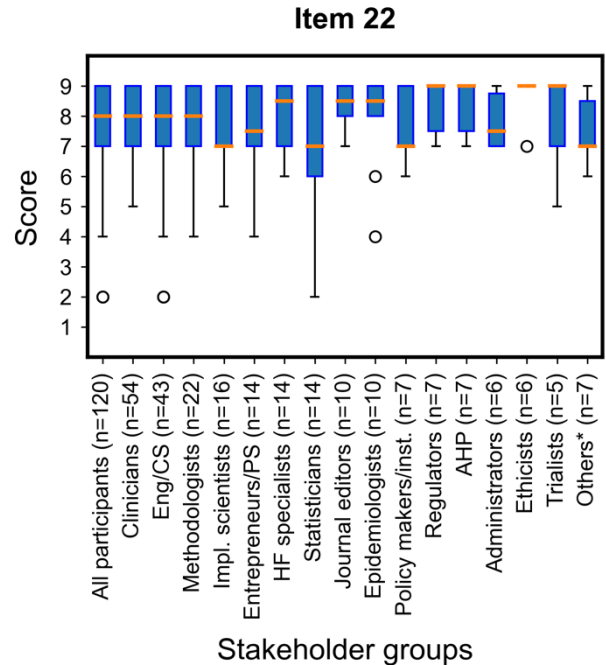


Figure 22: median score and IQR for item 22 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.7	1.5	84.0	1.7	0.8	None

Action taken: becomes item 13a in the updated list.

Item 23

Describe how algorithm recommendation/output errors were defined and how they were identified.

Number of comments: 3

Summarised participant comments:

- Include error variation between subgroups
- Distinguish between acceptable and unacceptable errors
- Could mention gold standard or ground truth used to identify errors?

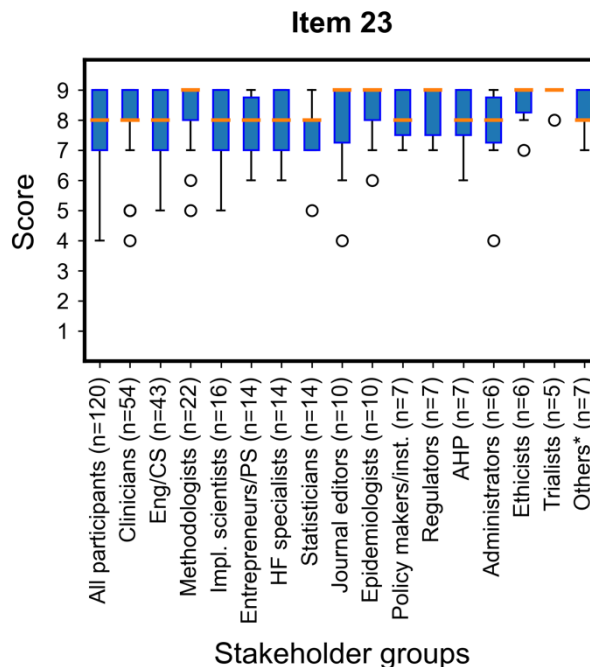


Figure 23: median score and IQR for item 23 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.9	1.2	90.0	0.0	0.0	None

Action taken: becomes item 11b in the updated list.

Item 24

Describe any attempts to familiarise users with the algorithm, including any training received.

Number of comments: 1

Summarised participant comments:

- Important to refer to any material used for training (allow reproducibility).

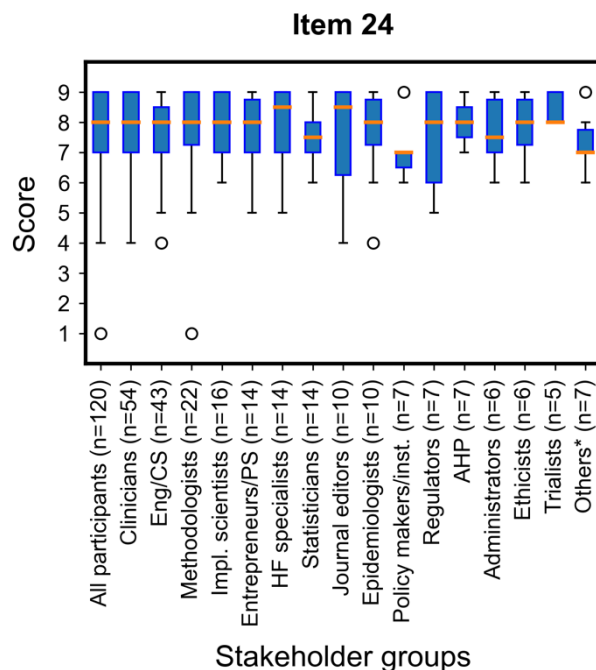


Figure 24: median score and IQR for item 24 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.5	1.4	81.5	0.8	0.8	None

Action taken: becomes item 8c in the updated list.

Item 25

Describe the human factors tools, methods or frameworks used to evaluate usability, situation awareness and any other relevant human factors considerations. Justify this choice.

Number of comments: 3

Summarised participant comments:

- Take out '*human factors*' as researchers may use implementation science or other methods instead
- More qualitative research is important to evaluate the quality of user interaction
- Important but not as much as other items.

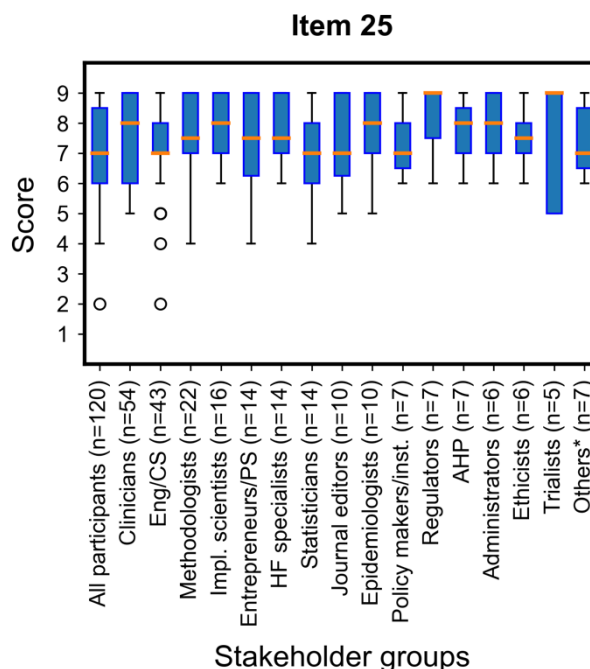


Figure 25: median score and IQR for item 25 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
7	6	8.5	7.3	1.4	70.6	0.8	0.8	Regulators, Trialists

Action taken: wording modified, merged with item 26 and 38 of the original list, becomes item 14 in the updated list.

Item 26

Describe any attempt to understand the user acceptance of the algorithm as well as user deviations from the algorithm's recommendations or intended use.

Number of comments: 5

Summarised participant comments:

- Two different things, should be split
- Better to assess acceptance in large trial, should rather look at deployment feasibility here
- User acceptance is important but not necessarily indicative of outcome (users can hate an algorithm but be forced to use it)
- User acceptance is vague
- User deviations critical to assess algorithm failure/patient risk.

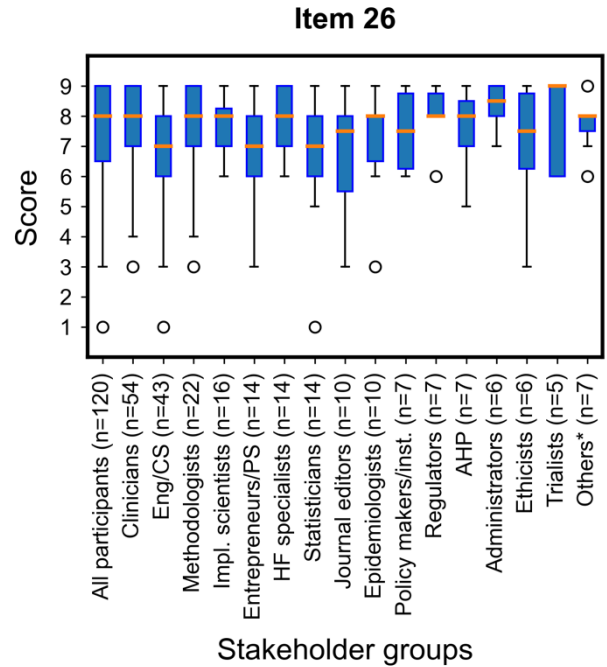


Figure 26: median score and IQR for item 26 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	6.5	9	7.3	1.7	74.8	4.2	0.8	None

Action taken: wording modified, merged with item 25 and 38 of the original list, becomes item 14 in the updated list.

Item 27

Describe any involvement of patients in understanding their opinion on the algorithm and how the algorithm's outputs could influence their care.

Number of comments: 4

Summarised participant comments:

- Important but often could report in a stand-alone report so shouldn't mandate them here
- Some algorithms may not affect patient care
- Mention how patient feedback was collected and incorporated.

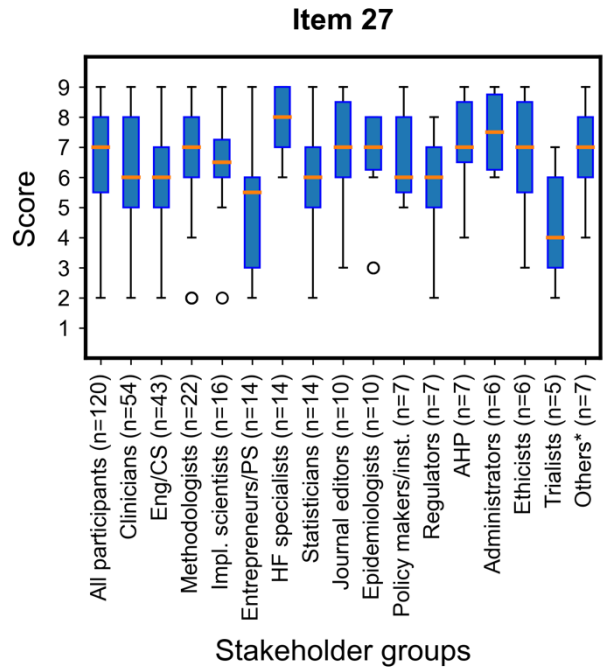


Figure 27: median score and IQR for item 27 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
7	5.5	8	6.5	2.0	54.6	10.1	0.8	Trialists

Action taken: wording and focus modified, merged with item 43 of the original list, becomes item 15 in the updated list.

Item 28

Describe the methodology used to collect and analyse data for the health economic assessment of the algorithm's use.

Number of comments: 12

Summarised participant comments:

- Important but often could be reported in a stand-alone report so shouldn't be mandated here
- Unlikely this will be done at this stage
- Add 'if applicable'
- May already be an outcome if the aim of the algorithm is to reduce costs
- A threshold analysis on how good the performance of the AI intervention and how large the cost of the intervention have to be to fall into a cost effective range could be more appropriate.

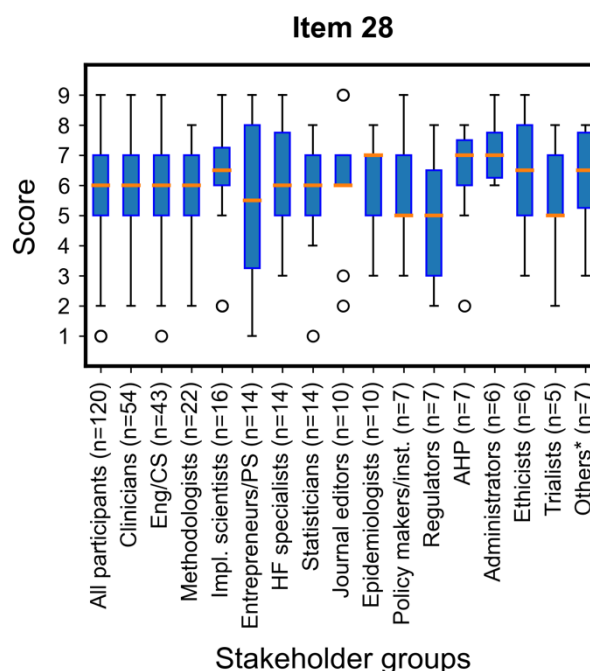


Figure 28: median score and IQR for item 28 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
6	5	7	6.0	1.9	42.7	12.0	2.6	None

Action taken: item dropped due to low consensus in the scoring exercise and congruent comments that health economic evaluation would not fit well within the scope of DECIDE-AI.

RESULTS

Item 29

Describe the user population baseline characteristics (number, number of centres, specialty, seniority, previous experience with digital support, etc.).

Number of comments: 3

Summarised participant comments:

- Merge with 30
- Overlap with 38
- make a distinction between the "study group" and the "study population" - the study group of participants will typically be a sample from the total population.

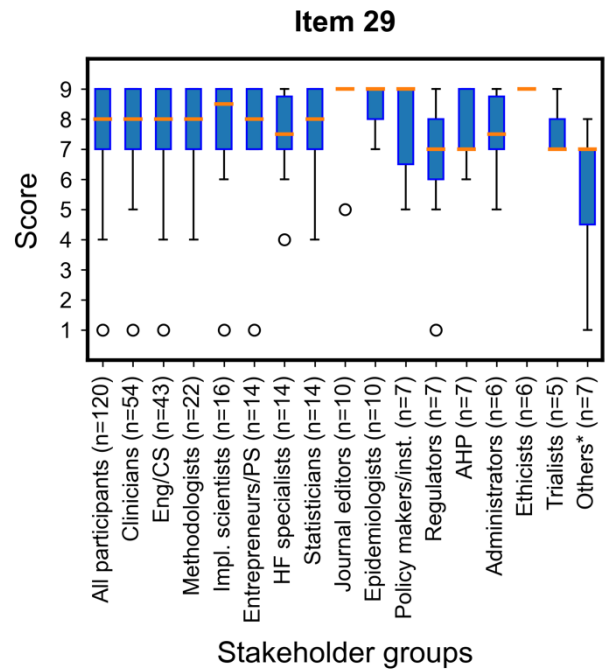


Figure 29: median score and IQR for item 29 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.9	1.4	87.5	0.8	0.0	None

Action taken: becomes item 17b in the updated list.

Item 30

Describe the patient population baseline characteristics (number, number of centres, age, sex, ethnicity if relevant, comorbidities, prevalence of the target conditions, etc.).

Number of comments: 2

Summarised participant comments:

- Merge with 29
- Please, make a distinction between the "study group" and the "study population" - the study group of participants will typically be a sample from the total population.

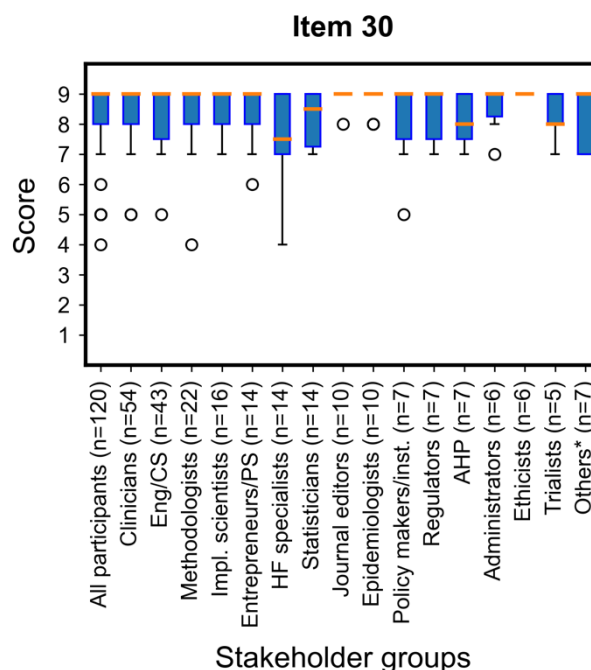


Figure 30: median score and IQR for item 30 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
9	8	9	8.3	1.0	95.8	0.0	0.0	None

Action taken: becomes item 17a in the updated list.

Item 31

Report on the proportion of intended users who had exposure to the algorithm (implementation reach), on the number of instances the algorithm was used (implementation dose) and on the users' compliance with the intended implementation (implementation fidelity).

Number of comments: 5

Summarised participant comments:

- Could add time spent with algorithm
- Not clear
- Implementation fidelity is a key consideration
- Describe how and by whom this was evaluated and any strategy used to improve fidelity
- Could change during the study.

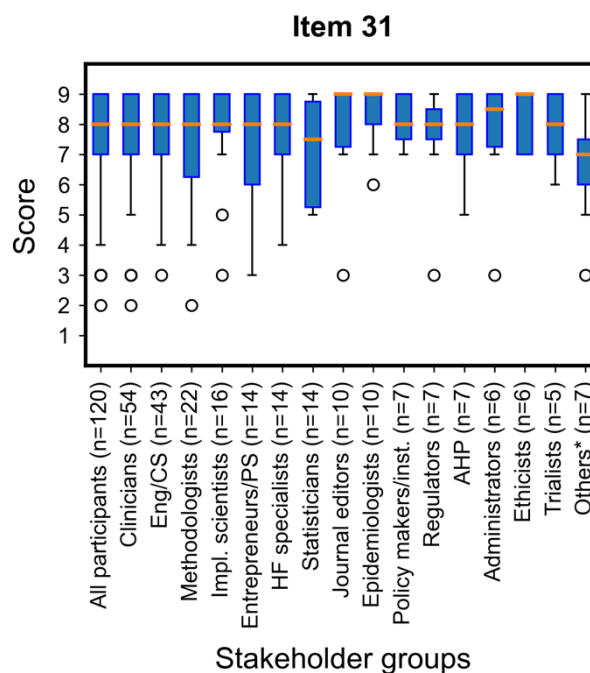


Figure 31: median score and IQR for item 31 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	7.5	1.6	81.4	3.4	1.7	None

Action taken: wording modified, becomes item 18a in the updated list.

Item 32

Report on the prespecified outcomes for the algorithm-assisted users (both overall and at an individual user level) as well as for the control group.

Number of comments: 4

Summarised participant comments:

- Not clear
- May not need a control group
- Should include an overall key performance metrics in the context of the clinical problem addressed.

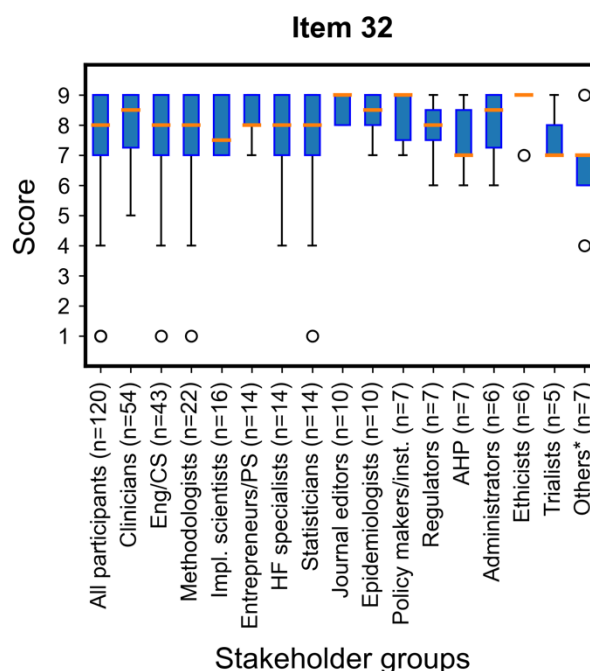


Figure 32: median score and IQR for item 32 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	7.9	1.5	89.0	1.7	1.7	None

Action taken: wording modified, split into item 20a and 20c in the updated list.

Item 33

If applicable, report on the prespecified outcomes which would have been observed had all the algorithm's recommendations been strictly followed.

Number of comments: 6

Summarised participant comments:

- Not clear
- Important even if not the main focus, as it could highlight the value of the system over and above the current clinical acceptance
- What's the point if the system is a decision support tool -> algorithm only is not the use case
- Should only use intention to treat, otherwise too many assumptions
- Counterfactual outcomes not likely to be available.

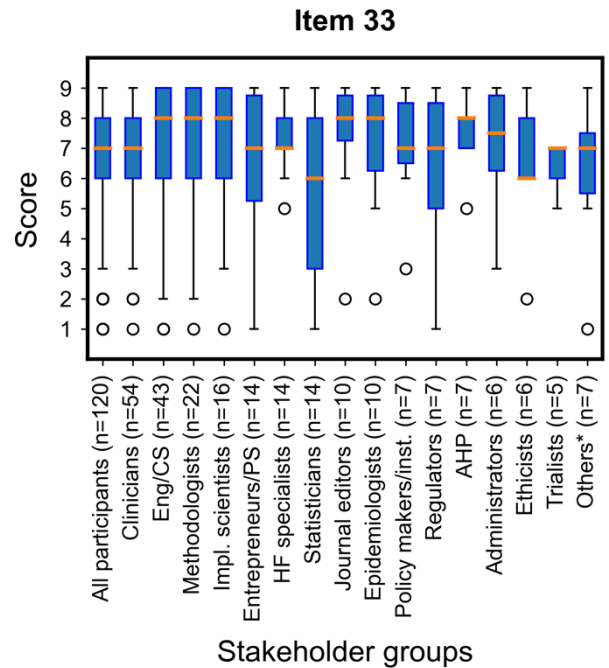


Figure 33: median score and IQR for item 33 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
7	6	8	6.9	2.1	66.4	10.3	3.4	None

Action taken: wording modified, becomes item 20b in the updated list.

Item 34

Describe any instances where the algorithm gave an erroneous recommendation/output. Report their rate of occurrence and detail their potential impact on patient care.

Number of comments: 7

Summarised participant comments:

- May not be known
- Not practical to report every instance depending on study size
- Erroneous might not be the correct term
- Error difference between subgroup is important.

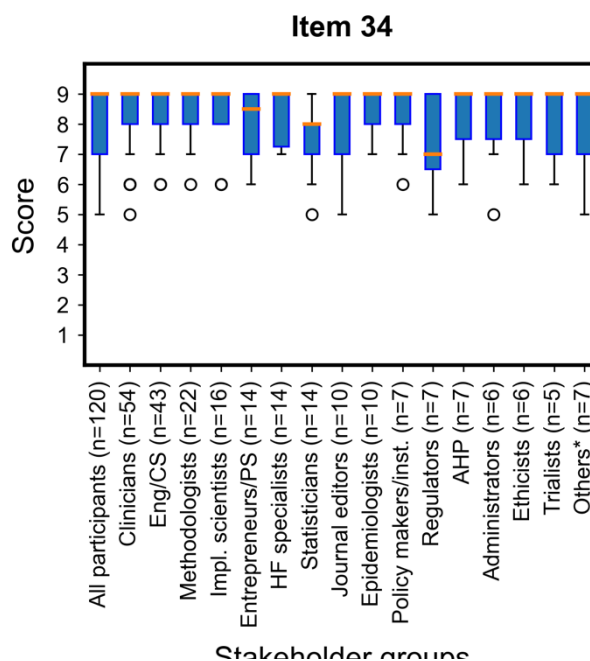


Figure 34: median score and IQR for item 34 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
9	7	9	8.2	1.1	90.7	0.0	1.7	Regulators

Action taken: wording modified, becomes item 21d in the updated list.

Item 35

Report on the compliance with the safety requirements.

Number of comments: 2

Summarised participant comments:

- Vague

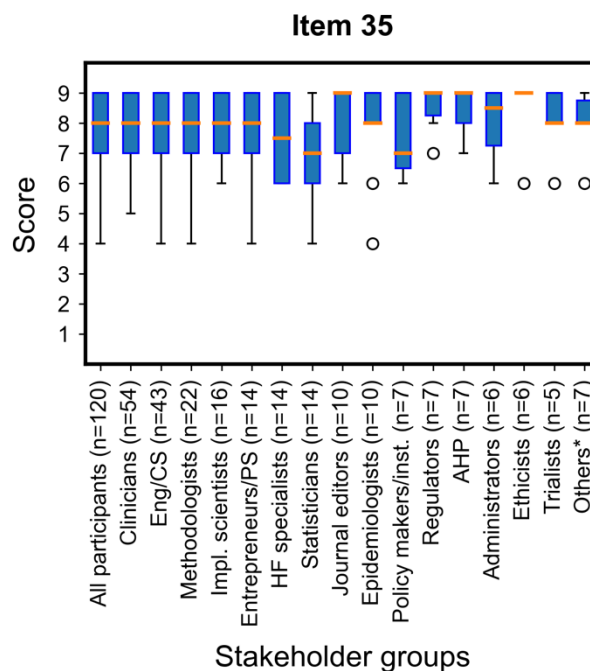


Figure 35: median score and IQR for item 35 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.8	1.3	82.1	0.0	2.6	None

Action taken: becomes item 21a in the updated list.

Item 36

Describe any instances where users decided to override the algorithm's recommendation or to follow an erroneous recommendation.

Number of comments: 5

Summarised participant comments:

- Not practical to report every instance depending on study size
- Assumes a gold standard but there may not be one
- Conflates two different things: automation bias vs clinician confidence/autonomy
- Needs reasons why overridden.

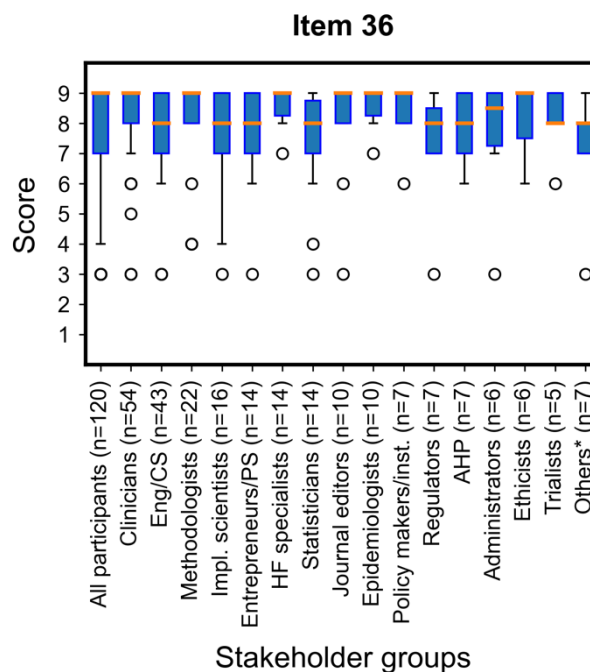


Figure 36: median score and IQR for item 36 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
9	7	9	8.0	1.5	88.3	3.3	0.0	None

Action taken: wording modified, becomes item 23a in the updated list, point about following erroneous recommendations moved to item 21d's explanation.

Item 37

Report on the evolution of users' trust in the algorithm (evolution of the overrides of the algorithm's recommendations with time) and on the learning curves (evolution of the users' performance with time).

Number of comments: 8

Summarised participant comments:

- Important but may be affected by information gained during the study
- May be under-powered at this stage
- People should have been fully trained before starting so ideally there would be no change in trust or learning curves
- Define trust or only focus on the evolution of overrides
- Learning when to trust the algorithm is a learning curve on its own
- Trust and overrides may not necessarily be correlated
- Excellent practice but may be difficult to assess in early stage study

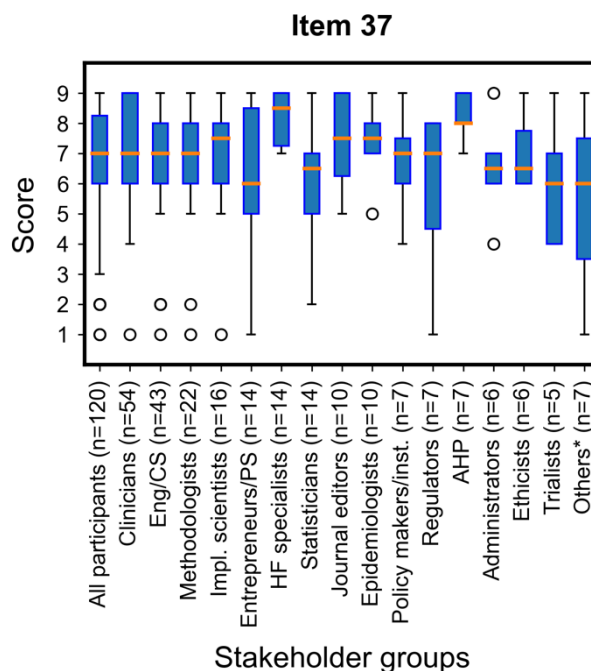


Figure 37: median score and IQR for item 37 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
7	6	8.25	7.0	1.8	65.0	5.0	0.0	None

Action taken: wording modified (brackets moved to explanation), becomes item 23b in the updated list, point about learning curves moved to item 23d of the updated list.

Item 38

Report the number of users involved in the human factors evaluation, their characteristics and the use cases examined.

Number of comments: 2

Summarised participant comments:

- Overlap with 29
- Should focus on decision making itself rather than just subjective views of AI.

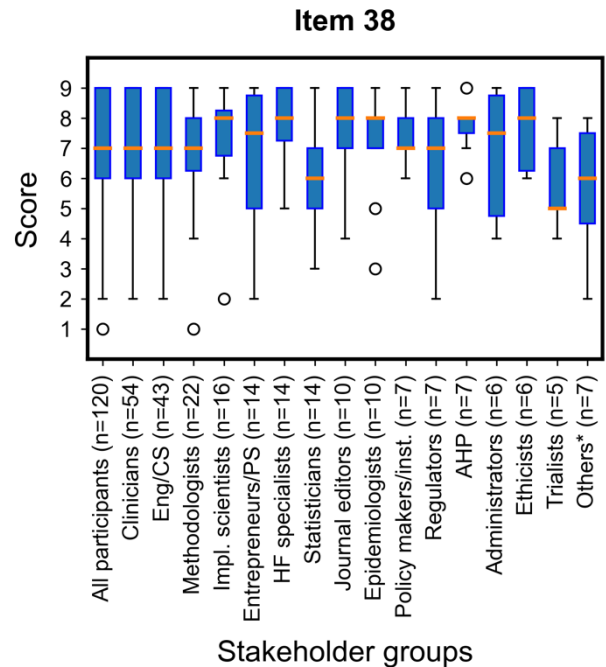


Figure 38: median score and IQR for item 38 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	6	9	7.1	1.8	66.9	5.1	1.7	Trialists

Action taken: wording modified, merged with item 25 and 26 of the original list, becomes item 14 in the updated list.

Item 39

Report on the usability evaluation, including time to task completion and display interface evaluation, using method-specific metrics.

Number of comments: 1

Summarised participant comments:

- Should add, *if applicable*

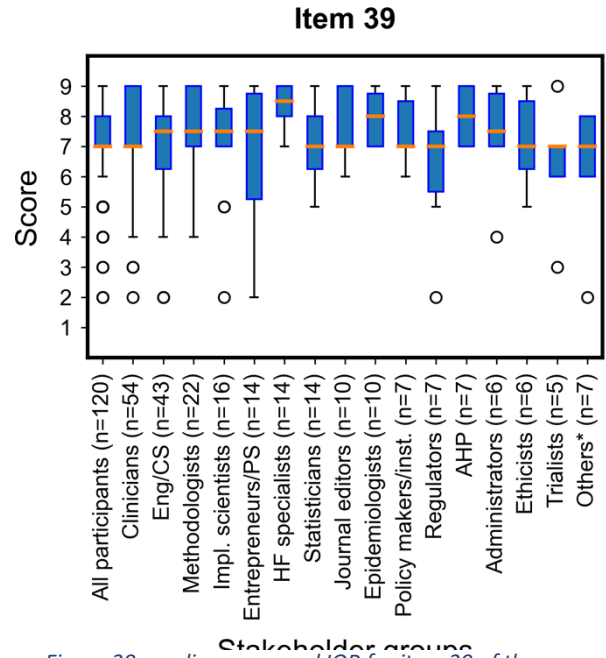


Figure 39: median score and IQR for item 39 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	7	8	7.3	1.6	75.6	3.4	0.8	None

Action taken: wording modified, becomes item 23c in the updated list.

Item 40

Report on the situation awareness evaluation and on the users' perspective on the algorithm's interpretability.

Number of comments: 5

Summarised participant comments:

- Merge with 41
- Should not be mandated
- Define interpretable: does it refer to explainable AI or user knowledge of what they were supposed to do in the study?
- Leave out situational awareness as this is an odd use of the term and focus the item on interpretability.

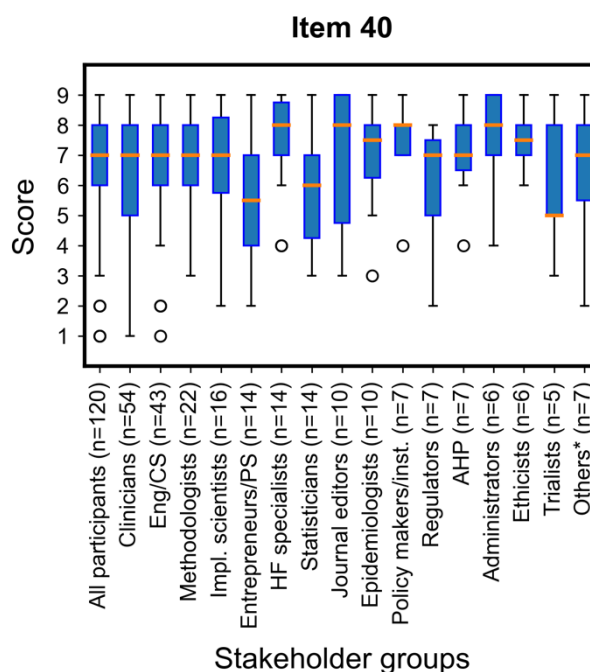


Figure 40: median score and IQR for item 40 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
7	6	8	6.7	1.9	64.7	7.6	0.8	Trialists

Action taken: wording modified, becomes item 23e in the updated list.

Item 41

Report on the outcomes of any other human factors evaluation, including the user acceptance of the algorithm and any induced changes in the care pathway.

Number of comments: 3

Summarised participant comments:

- Merge with 40
- Edit to '*the algorithm and its explanations*'.

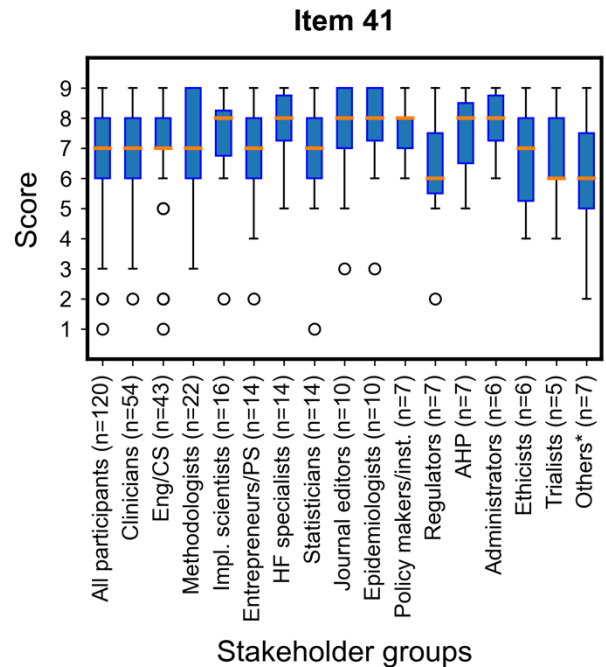


Figure 41: median score and IQR for item 41 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	6	8	7.0	1.7	70.3	5.1	1.7	None

Action taken: wording modified, becomes item 18b in the updated list, point about acceptance moved to item 23c of the updated list.

Item 42

Summarize all 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 design-evaluation cycle.

Number of comments: 7

Summarised participant comments:

- Should not tacitly recommend changing algorithm during course of a small study, leave that to earlier development phases
- Don't need too much details, just if changes happened, what changes these were and if they had an impact on the results
- Shouldn't need to be specified unless it has an impact
- Safety/efficacy cannot be evaluated during a rapid cyclical mod-eval series. Choose one or the other
- This is important and can be seen as the algorithm/system learning curve.

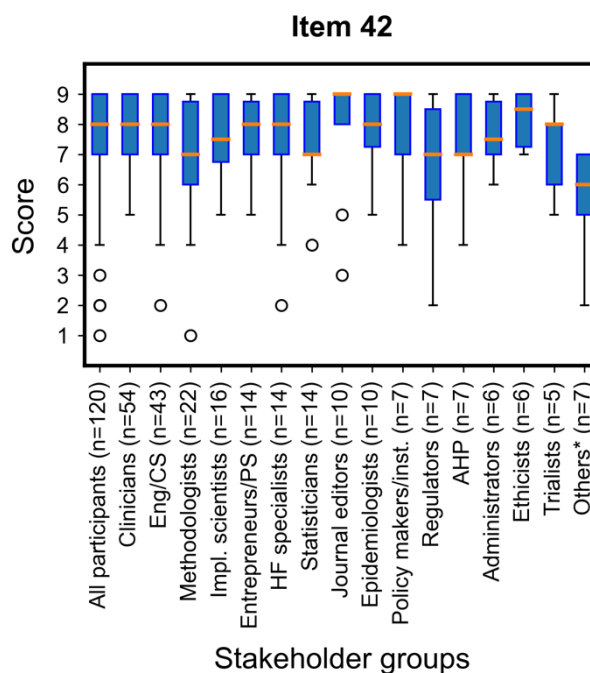


Figure 42: median score and IQR for item 42 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.5	1.7	78.3	3.3	0.0	Others

Action taken: wording modified, becomes item 19 in the updated list.

Item 43

Report the patients' opinion on the algorithm and whether they would accept their care being influenced by it.

Number of comments: 11

Summarised participant comments:

- May not be patient facing algorithm
- Valuable but better done/reported in separate research
- Patients might not have the requisite knowledge to appraise the algorithm itself. This type of evaluation may be less robust than imagined
- Better to solicit patient feedback into the process rather than algorithm itself per se
- Add 'if applicable'
- Won't patient's already have consented to the algorithm influencing their care?
- More colour should be added to the items on patient and public engagement to establish what level of analysis are deemed essential.

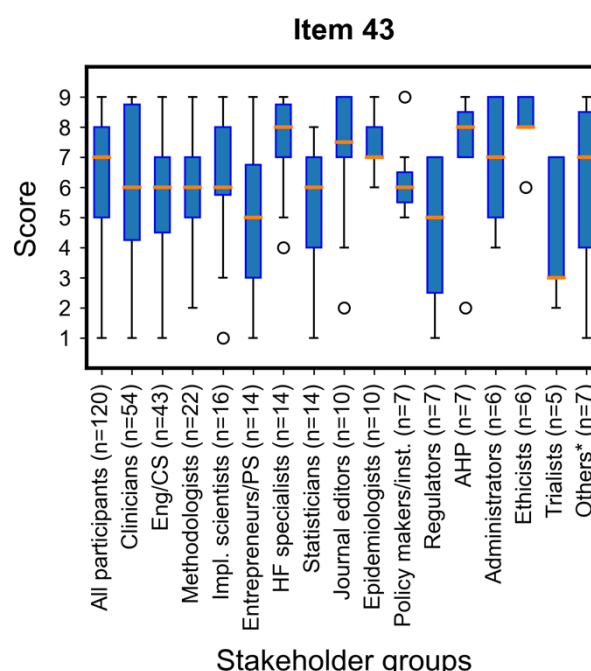


Figure 43: median score and IQR for item 43 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	5	8	6.4	2.2	55.9	12.7	1.7	Entrepreneurs/private sector Regulators Trialists

Action taken: wording and focus modified, merged with item 27 of the original list, becomes item 15 of the updated list.

Item 44

Report the results of the health economic assessment of the algorithm's use and identify any trade-offs in the care pathway.

Number of comments: 15

Summarised participant comments:

- Improve phrasing
- Valuable but better done/reported in separate research or at later stage during main big trial
- an understanding and appreciation of pathway changes and potential value is really helpful at this stage and important to capture
- More colour should be added to the items on economic outcomes assessment to establish what level of analysis are deemed essential.

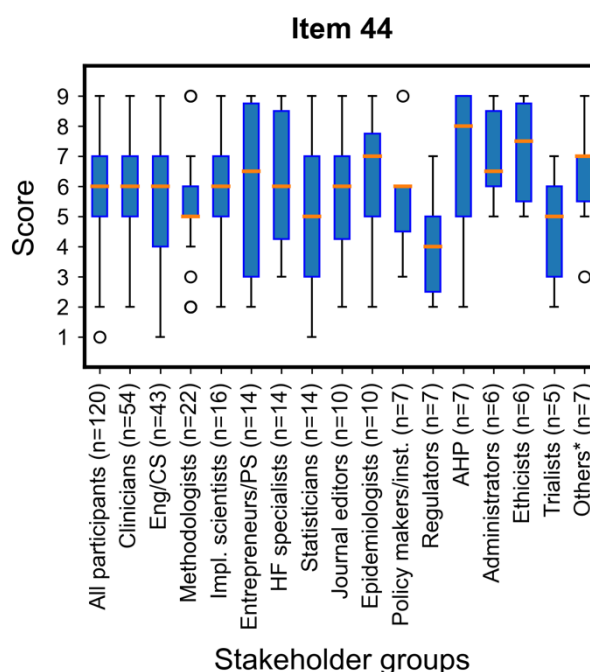


Figure 44: median score and IQR for item 44 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
6	5	7	5.8	2.1	37.0	15.1	0.8	Regulators Allied health professionals

Action taken: item dropped due to low consensus in the scoring exercise and congruent comments that health economic evaluation would not fit well within the scope of DECIDE-AI.

DISCUSSION

Item 45

Discuss if the obtained results support the intended purpose of the algorithm in real world healthcare settings, including how the outcomes would translate into patient benefit, or if an alternative use could be more appropriate.

Number of comments: 3

Summarised participant comments:

- Intended purpose fine but patient benefit speculative at this early stage. A focus on the latter should not be encouraged
- Should include a discussion on the benchmark aim for the overall performance and how this was arrived at.

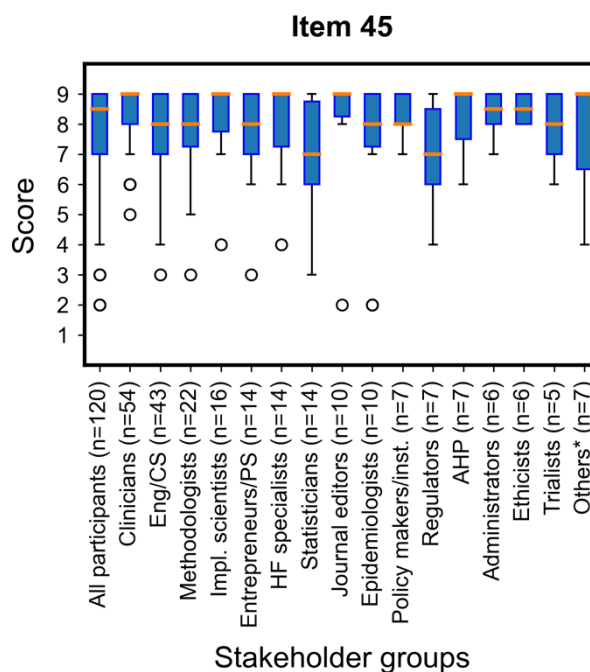


Figure 45: median score and IQR for item 45 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8.5	7	9	7.9	1.6	86.7	3.3	0.0	None

Action taken: wording modified, becomes item 24 in the updated list.

Item 46

Explain what was learned about the reasons for human deviation from the algorithm's recommendations or intended use, and what this tells us about achieving better alignment.

Number of comments: 1

Summarised participant comments:

- deviations may be equally legitimate choices so alignment may be a misnomer.

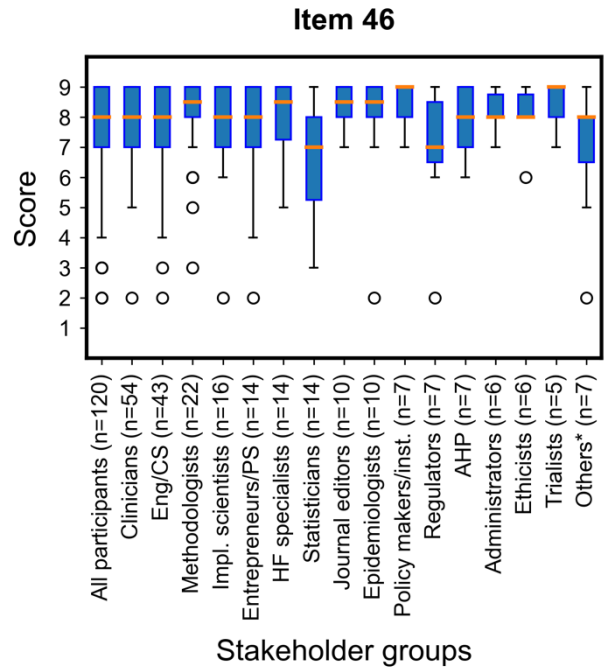


Figure 46: median score and IQR for item 46 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.7	1.5	85.8	3.3	0.0	None

Action taken: wording modified, merged with item 49 of the original list, becomes item 26 in the updated list.

Item 47

Discuss the algorithm's errors and identify any underlying pattern or algorithmic bias. Explain how these can be mitigated.

Number of comments: 6

Summarised participant comments:

- Define algorithmic bias
- Algorithmic bias can be challenging to identify and mitigate
- if there are few enough examples to report every error, it will be hard to draw any underlying patterns from it. If there are enough examples, then reporting on every error would not be appropriate
- Is an algorithm error a 'fault' or a recommendation that turned out not to be good? Clarify. Latter can be hard/impossible to define
- Discuss how errors of the system relate to the frequency and severity of human errors (It should be about the benefit risk ratio, not about errors perse).

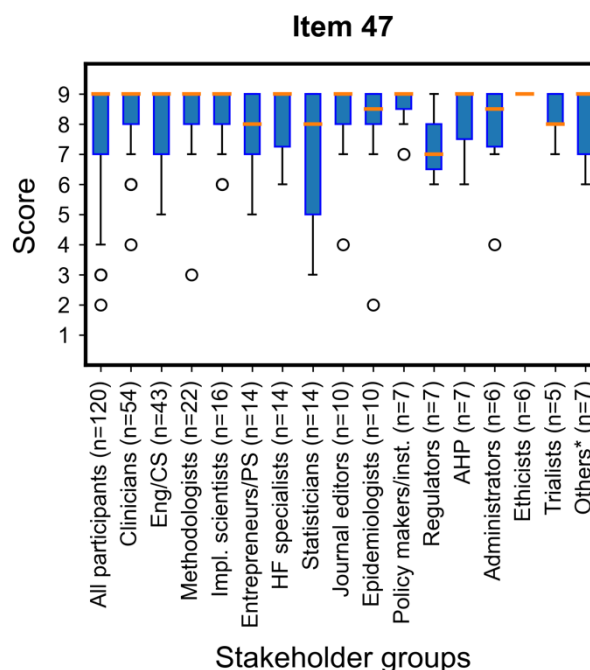


Figure 47: median score and IQR for item 47 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
9	7	9	8.0	1.4	89.1	2.5	0.8	Regulators

Action taken: wording modified, merged with item 48 of the original list, becomes item 25 in the updated list.

Item 48

Discuss what the results suggest about the safety profile of the algorithm.

Number of comments: 1

Summarised participant comments:

- should include demonstration of conformity to ISO 62366 (mandatory regulatory process for Class II and above devices).

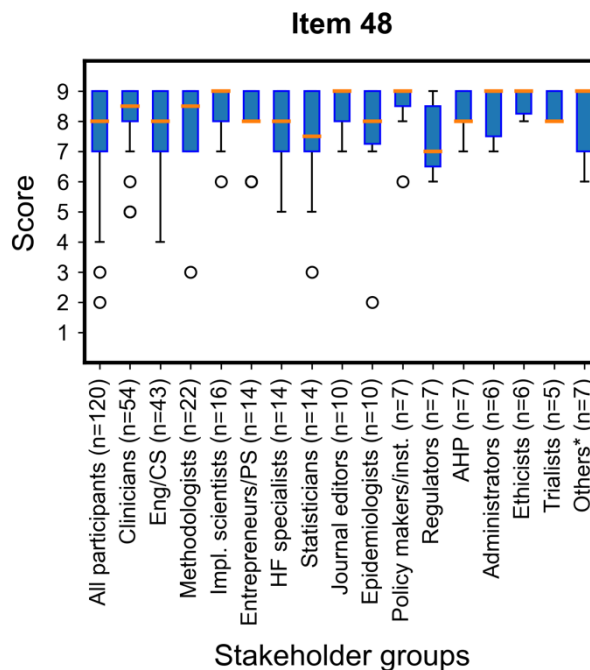


Figure 48: median score and IQR for item 48 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	8.0	1.3	90.0	1.7	0.0	None

Action taken: merged with item 47 of the original list, becomes item 25 in the updated list.

Item 49

Discuss the human factors results and comment on the evolution of the algorithm/hardware platform design. Discuss the need for additional technical requirements or product design improvement before large-scale summative evaluation.

Number of comments: 3

Summarised participant comments:

- Remove human factors from the item description
- Unclear why this clinical study has design-evaluation iteration, presumed that was already done in development phase.

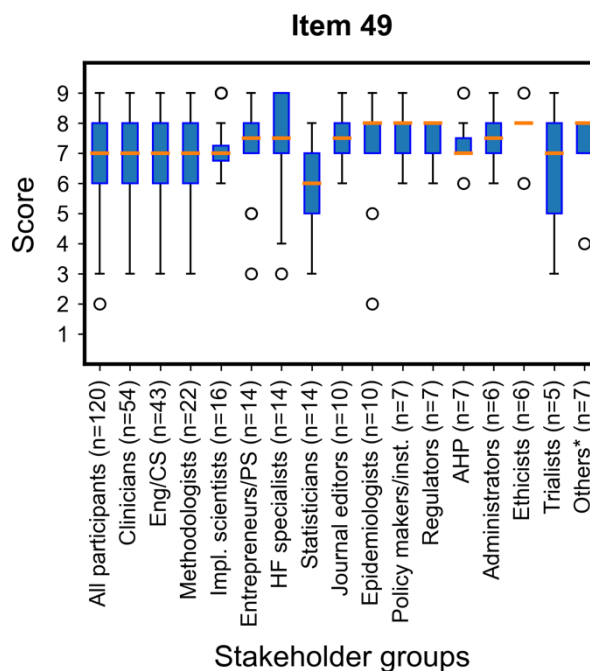


Figure 49: median score and IQR for item 49 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7	6	8	7.0	1.6	70.9	4.3	2.6	None

Action taken: wording modified, merged with item 46 of the original list, becomes item 26 in the updated list, point about additional technical requirements and product design moved to item 27 in the updated list.

Item 50

Comment on the evolution of users' trust in the algorithm and on the learning curves. State when they reached a stable state.

Number of comments: 9

Summarised participant comments:

- Conflating two potentially different issues
- evolution of trust and evolution of human performance is a failure of study design and user training in many circumstances so the wording should be changed
- Might be beyond the scope of a small study, may not reach steady state with small sample
- Trust might not be best word (trust is relational, between humans), instead use confidence with algorithm
- It is vital to explore trust and confidence
- Alignment of trust with trustworthiness is key (Do users notice when the algorithm has reached the limits of the performance window and react appropriately?)
- Why even measure trust? This is not done for other devices, and this will overburden AI researchers.

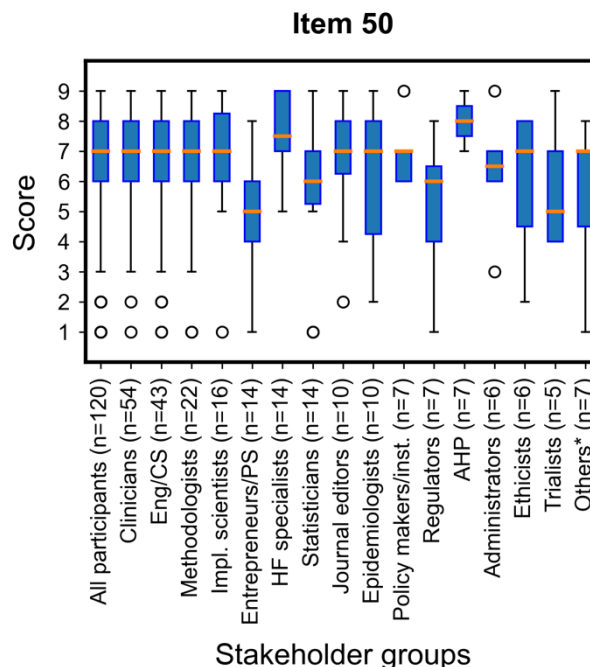


Figure 50: median score and IQR for item 50 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=8 points away from overall median
7	6	8	6.6	1.9	59.7	7.6	0.8	Entrepreneurs/private sector Trialists

Action taken: merged with item 51 and 52 of the original list, becomes item 27 in the updated list, most content moved to the item's explanation.

Item 51

Highlight any performance difference in user or patient subgroups and discuss the merits of limiting further evaluation to a specific group of users or patients.

Number of comments: 5

Summarised participant comments:

- Clinical trials are already often too narrow in their inclusion criteria, why should limitation to specific subgroups be a desirable outcome (specially in the context of black boxes and related generalization understanding issues)?
- Algorithms should be altered at the development stage if it doesn't generalize well, not at the trial stage
- De-prioritize this as sample size will be too small to reliably tell
- this should be used to determine the parameters or required data collection in a larger study
- should discuss what was done to actively reduce variation in subgroup performance.

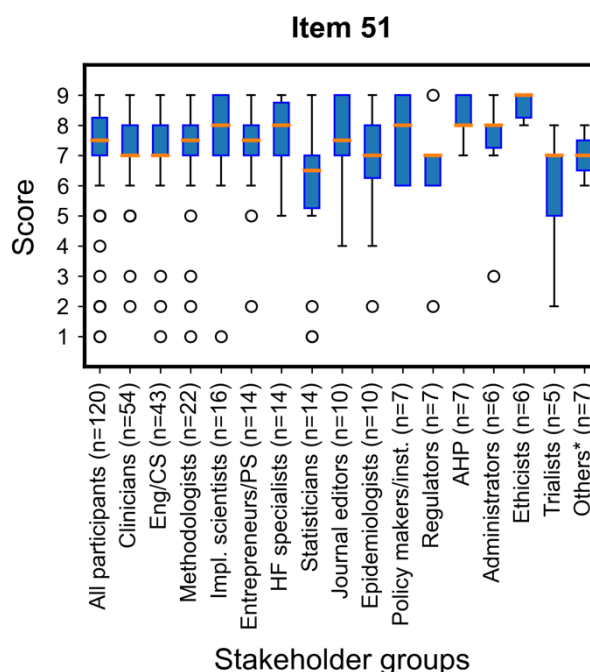


Figure 51: median score and IQR for item 51 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
7.5	7	8.25	7.2	1.8	78.3	5.8	0.0	None

Action taken: merged with item 50 and 52 of the original list, becomes item 27 in the updated list, most content moved to the item's explanation.

Item 52

Discuss the feasibility and appropriateness of large-scale summative evaluation in light of the obtained results.

Number of comments: 2

Summarised participant comments:

- unclear wording
- Not just evaluation but also feasibility of large-scale deployment at all (e.g. too resource intensive etc.).

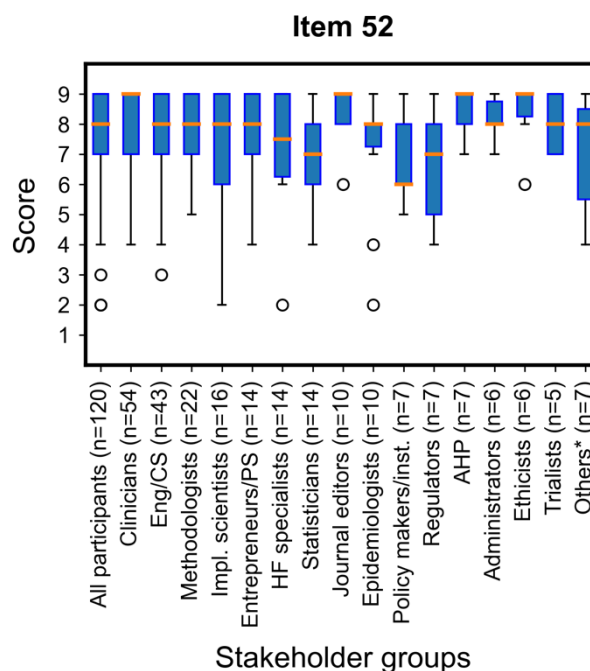


Figure 52: median score and IQR for item 52 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring >=7	% scoring <=3	% "I don't know"	Stakeholder group with median <=2 or >=2 points away from overall median
8	7	9	7.8	1.5	82.4	2.5	0.8	Policy makers

Action taken: merged with item 50 and 51 of the original list, becomes item 27 in the updated list, most content moved to the item's explanation.

STATEMENTS

Item 53

Disclose the source of funding for the study and authors' relevant conflicts of interest.

Number of comments: 6

Summarised participant comments:

- Role of funders as well as source of funding should be disclosed
- Disclose individual contributions to the study
- so obvious that it does not need to be included
- conflicts of interest are to be expected and it is crucial that there is transparency in reporting of the study
- include any type of involvement of commercial companies in the study design or study itself.

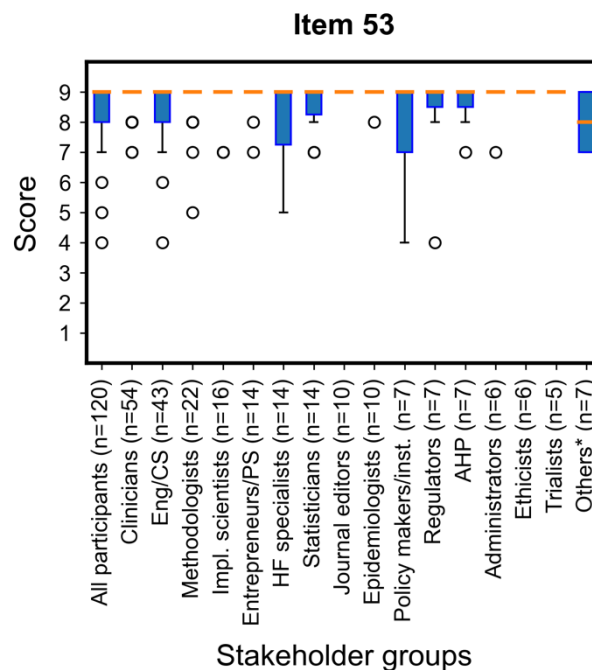


Figure 53: median score and IQR for item 53 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	
9	8	9	8.5	0.9	97.5	0.0	0.0	

Action taken: wording modified, becomes item 29 in the updated list.

Item 54

Disclose code and data availability.

Number of comments: 13

Summarised participant comments:

- Should make clear that this is about *if* they are available rather than stating that they *should* be available
- Level of availability should also be mentioned
- State *how* data could be made available to authorized/bona fide researchers
- Ideal but not realistic, might deter tech companies, could be proprietary, data security and anonymization issues
- Shouldn't be asked
- The sticks to enforce this probably sit better with regulators than voluntary publishing standards
- Explain lack of availability if that's the case and explain how the veracity of the model is shown if the data and code have not been peer-reviewed
- it would be useful to recommend standard statements where data availability is restricted by law
- when it is about availability, it only makes sense in non-commercial applications, otherwise it becomes an advertisement
- should be reworded *Disclose code and data availability where possible and whether collaboration possible.*

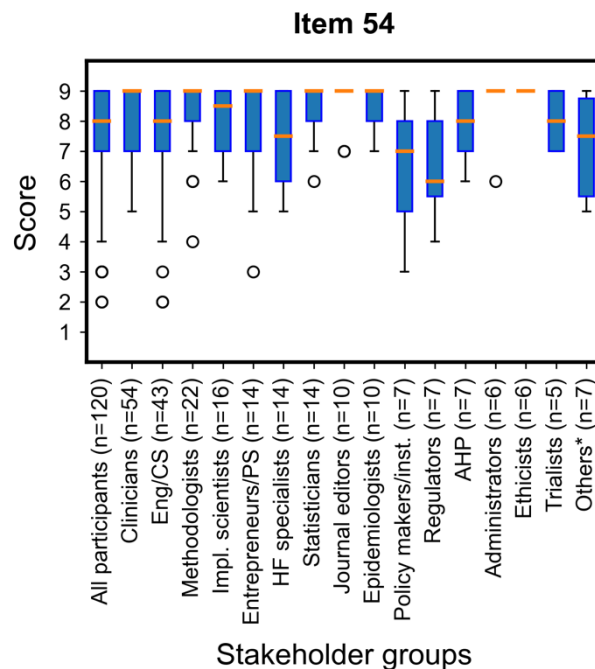


Figure 54: median score and IQR for item 54 of the original list, by stakeholder groups. Whiskers represent the last value comprised within 1.5 IQR on both side of the median. Eng = Engineers, CS = Computer Scientists, PS = Private sector, HF = Human factors, AHP = Allied health professionals, *Patient representatives and psychologists.

Summary statistics of the scores obtained in Round 1 (all participants)								
Median	IQR-25	IQR-75	Mean	Standard deviation	% scoring ≥ 7	% scoring ≤ 3	% "I don't know"	Stakeholder group with median ≤ 2 or ≥ 2 points away from overall median
8	7	9	7.7	1.7	80.3	4.3	2.6	Regulators

Action taken: wording modified, becomes item 30 in the updated list.

SECTION SPECIFIC COMMENTS

INTRODUCTION

Summarised participant comments:

- Not necessary to go into lots of detail about development if previously published
- Need anticipated use and setting
- Include a clear statement of limitations
- Items good and critical to establish the justification for the AI and therefore the likelihood of meeting ethical requirements such as beneficence and justice.

METHOD

Summarised participant comments:

- A priori identification of any patient and user subgroups is necessary
- Specify if iterative modification of algorithm happened during study and if so, what impact it had
- Greater emphasis on qualitative research -> e.g. user interaction, human factors, design etc.
- Patient involvement and health economic work not necessary at this stage
- Software/hardware detail should be reported so reader knows if it would work in their hospital
- Data completeness, imputation and source all need to be mentioned
- Possibly too many study types being performed at once here
- Ethics issues extend beyond just patient privacy, authors should comment on them.

RESULTS

Summarised participant comments:

- Too many assumptions beyond methods section and on very specific methods
- More detail on patient/economics aspects are desirable
- Too many aspects for one paper
- The scope of DECIDE-AI needs to be more clearly defined
- some of these items would be expected in a protocol writing guide, not in a reporting guideline. The logic for including items should be guided by arguments that make it clear that absence of that information makes it difficult to appreciate the validity or applicability of the study findings.

DISCUSSION

Summarised participant comments:

- Not all necessary for an early phase report
- All important areas
- May be better in different papers rather than all in one report
- Too long
- Very context dependent
- Discussion section will not contain essential info if the rest is complete.