

DECIDE-AI

Consensus Group offline exchanges – item list

01.06.21

How to read this document?

For each of the item in the updated list, following information is provided from the Round 2 analysis.

1. the median score for the 136 participants
2. the percentage of participants who attributed a score greater or equal to 7; in other words, attributed a score meaning than the item should be included
3. the stakeholder groups whose median score was 2 points or more away (in positive or negative) from the overall median, if any
4. the number of comments made on the item specifically. This number is to be interpreted in consideration of the overall number of respondents (n=136)
5. a summary of the comments made. The summary was written in parallel by two members of the research team (Myura Nagendran and Baptiste Vasey) and conflict were resolved by consensus. No indication is given about the frequency of each individual comment.
6. a figure representing the different scores attributed graphically. Figure legend: median score and interquartile range (IQR), overall and by stakeholder group. Whiskers represent the last value comprised within 1.5 IQR on both side of the median and circles represent outliers. Eng = Engineers, CS = Computer Scientists, PS = Private Sector, HF = Human Factors, PM = Policy Makers AHP = Allied Health Professional, * Patients representatives, Funders and Psychologists. A same participant can be represented in several stakeholder groups (according to their own denomination).

TITLE AND ABSTRACT

Item 1a - Title/abstract

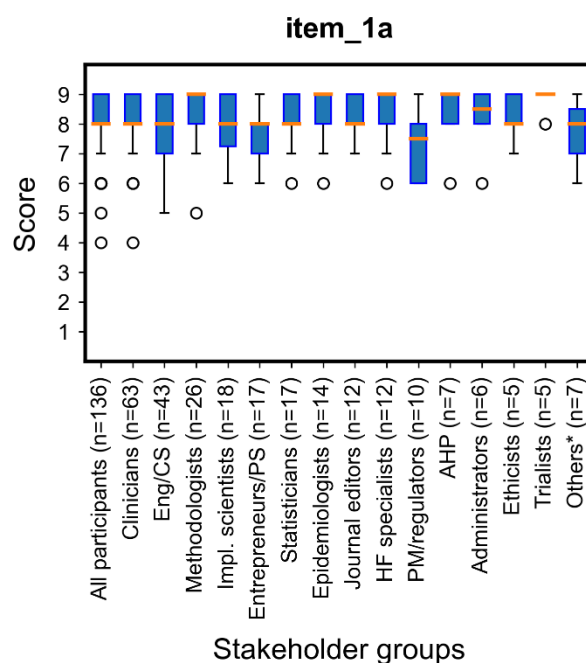
Identify the study as early stage or formative clinical evaluation of an artificial intelligence or machine learning based decision support system, mentioning the clinical problem addressed.

Overall median: 8

93.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 22

Summary of comments:

- Add some idea of study design
- Settle on either 'formative' or 'early-stage', align community around a single term if possible, perhaps with more elaboration in the E&E document
- Scope and potential overlap with TRIPOD -> consider mentioning that it is 'live' evaluation
- Scope implied by the word 'clinical' in clinical problem (all target problem might not be only clinical)
- Some want more detail on clinical problem, some say could be removed as will be in abstract
- Some reports still use "computerised clinical decision support system/tool" as denomination.

Figure 1. Median score and IQR for item 1a of the revised list, by stakeholder groups. Full legend on page 1.

Item 1b - Title/abstract

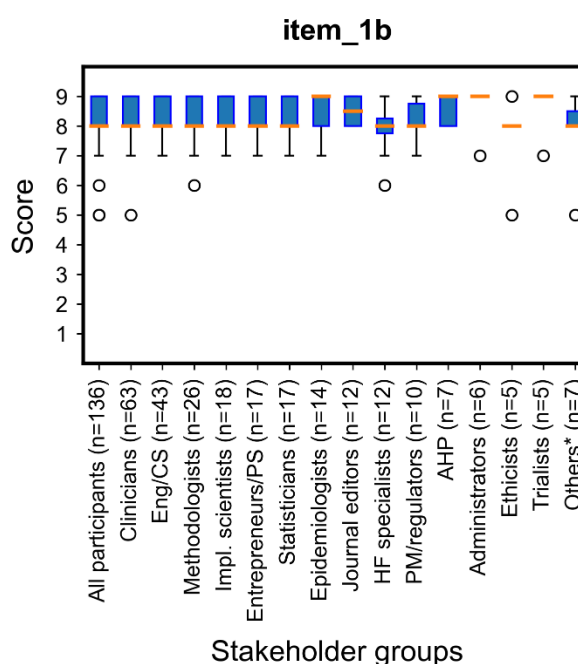
Provide a structured summary of the study, including: target clinical problem, intended use of the algorithm and integration in the clinical pathway, type of algorithm, study design, study setting, number of patients and users included, control group if applicable, primary and secondary outcomes, key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

Overall median: 8

97.8 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 31

Summary of comments:

- Specific guidance for the abstract should be developed
- Journal structure/word limits could affect reporting this item
- Could compress (e.g. human factors in results instead, algorithm type not as relevant in abstract, safety endpoints are outcomes, merge intended use and clinical integration)
- Overlap with item 1a
- Some participants want more detail on or mention of:
 - Users description to address e.g. bias issues
 - Data used to train/test the algorithm
 - Whether fixed or continuous learning
 - Whether regulatory approval granted or not
 - Gold standard/reference/definition for outcome
 - Software/hardware used
 - Study limitations.

Figure 2. Median score and IQR for item 1b of the revised list, by stakeholder groups. Full legend on page 1.

INTRODUCTION

Item 2 - Target clinical problem and population

Describe the target clinical problem and medical condition, including the current state of the art practice, and the target patient population.

Overall median: 8

96.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None

Number of comments: 14

Summary of comments:

- Few comments that the item could be compressed: clinical problem and population more important than 'state of the art'
- Few comments that the item is very important, one comment that it is redundant if already mentioned in pre-clinical paper
- Consider adding why researchers feel AI/ML approach was useful/necessary i.e. why *should* we rather than why *could* we
- 'State of the art' vague -> do you mean best clinical practice or best existing AI? -> could replace with 'current standard of practice' if the former to make it clearer.

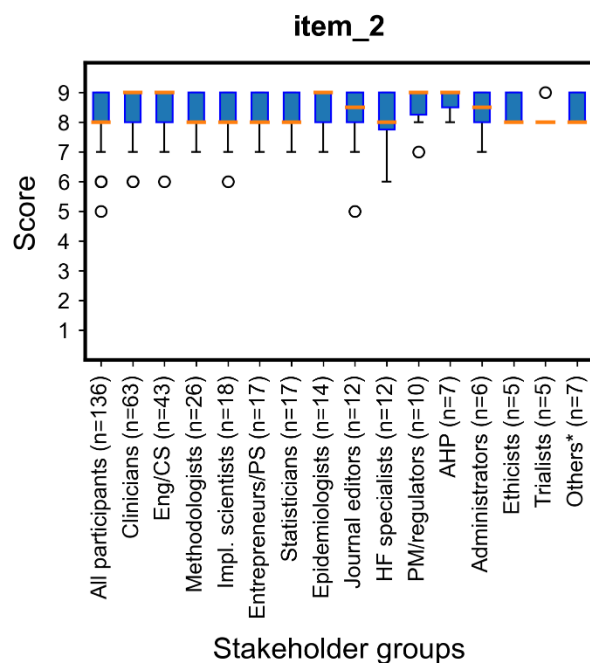


Figure 3. Median score and IQR for item 2 of the revised list, by stakeholder groups. Full legend on page 1.

Item 3 - Intended use

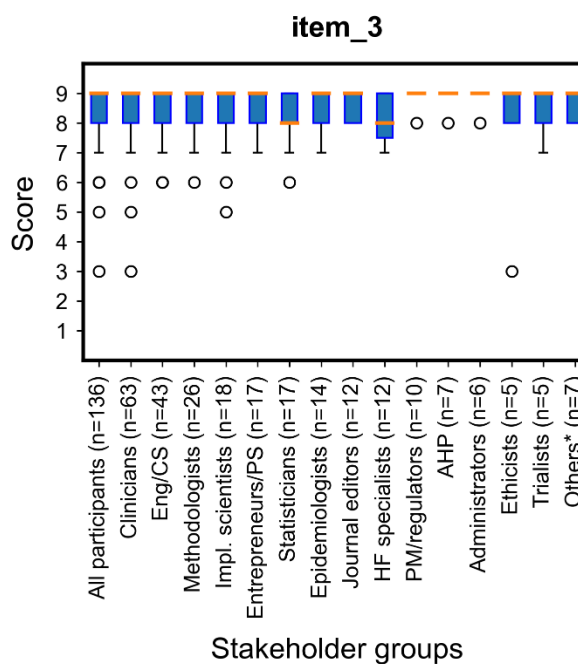
Describe the intended use of the algorithm, its planned integration in the care pathway and the impact in terms of patient outcomes it intends to achieve.

Overall median: 9

96.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 30

Summary of comments:

- Impact (including what i.e. savings/speed/clinical/patient experience) is more relevant here, could move integration to methods
 - Also consider renaming to 'potential' impact or 'objective'
- Stick to 'intended use' as more general term rather than the regulatory definition, more generally could consider box of key terms and definitions
- Could merge with item 2
- 'Patient' outcomes may be too narrow -> e.g. other benefits from speed/ease of use for clinician etc.
- Could move to conclusion/discussion, results may inform aspects of intended use etc.
- Consider flow chart/illustration
- regulatory bodies can - and often do - consider journal articles to go to the intended use of the device, depending on the framing. In short, regulatory bodies can often hold manufacturers accountable for the claims they make in these forums.

Item 4 - Current stage of development

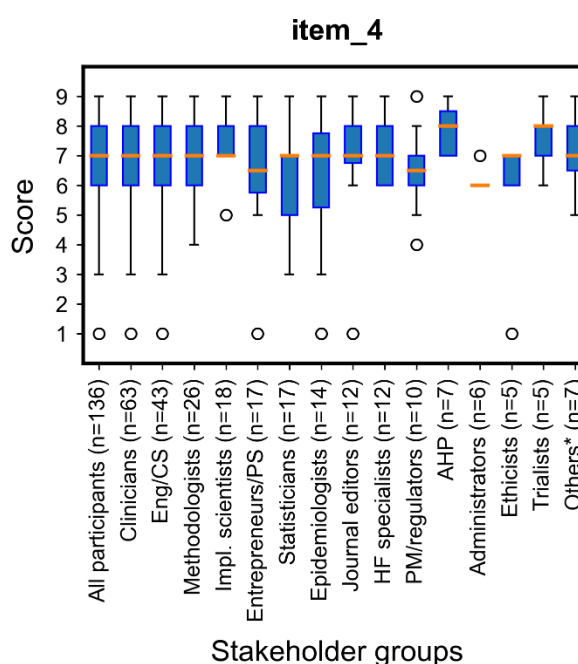
Describe the current stage of development of the algorithm (both from a scientific and a regulatory perspective). State if the algorithm is tested as a medical device and, if so, which regulatory approval is sought/was obtained.

Overall median: 7

72.4 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 26

Summary of comments:

Figure 5. Median score and IQR for item 4 of the revised list, by stakeholder groups. Full legend on page 1.

- Second sentence (regulatory aspect):
 - Might be better to leave scientific and regulatory evaluation separate
 - Results should be considered independently of regulatory status
 - Could be confidential if pending approval
 - May be too premature for regulatory items as frameworks still evolving
 - Might not be important in a clinical paper
 - Could move to methods
 - Importance of reporting regulatory aspects depends on how close to market, more important if closer
 - Unclear -> is this asking for approval or the study that led to regulatory approval?
- Could replace 'tested' with 'evaluated'
- Describe stage of development for each components (data collector, model, device, interface) separately
- Important to understand how far away a real life use of the algorithm is.
- Standardising terminology would be helpful.

Item 5 - Objectives

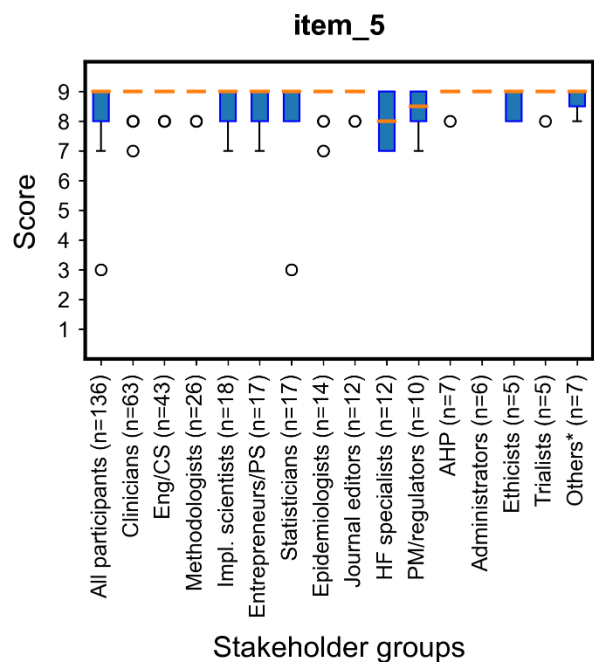
State the study objectives.

Overall median: 9

99.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 6

Summary of comments:

- Mixed. Some comments say redundant/overlaps with other items, some say important
- Very important as testing of these systems should be held to parallel standards to clinical trials
- Some comments that the item is too vague.

Figure 6. Median score and IQR for item 5 of the revised list, by stakeholder groups. Full legend on page 1.

METHODS

Item 6a - Research governance

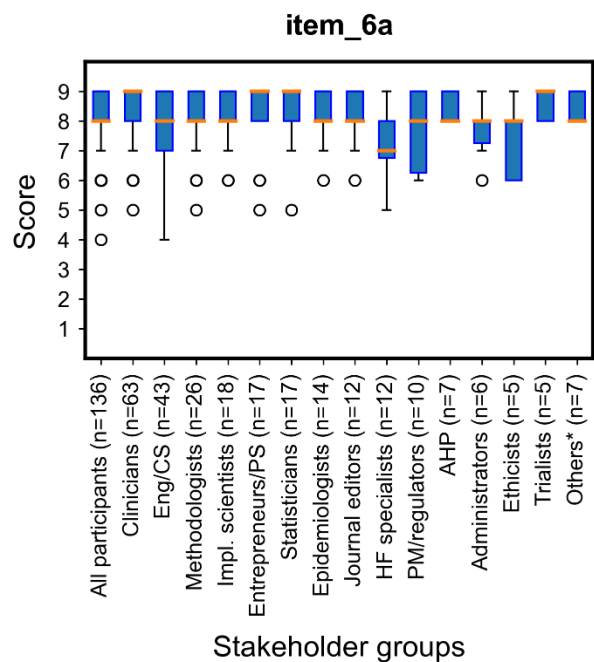
Provide a reference to any study protocol, study registration number and ethics approval.

Overall median: 8

89.6 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 13

Summary of comments:

- Could add 'or explain why it wasn't required / relevant'
- Some comments saying protocol not mandatory, so add 'if applicable'
- Some comments say pre-registration wouldn't be expected for early studies
- Could move regulatory points from item 5 here
- Lots of comments saying ethics would be mandatory/already part of good practice in many case.

Figure 7. Median score and IQR for item 6a of the revised list, by stakeholder groups. Full legend on page 1.

Item 6b - Research governance

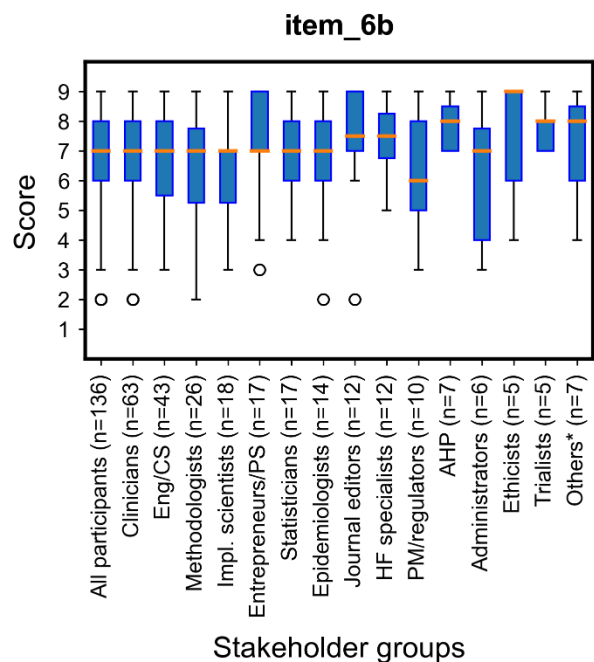
State what measures were taken to protect patient privacy and data security.

Overall median: 7

69.1 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- Ethicists



Number of comments: 23

Summary of comments:

- Most comments saying this may be redundant and could leave details in the study protocol instead (and/or include ethics/IRB application with protocol)
- Few comments stating the ethics is a governance/legal responsibility rather than something for reporting standards, is the responsibility for the IRB/ethics board
- Already covered by ethics, if any extra info will be in item 16.

Figure 8. Median score and IQR for item 6b of the revised list, by stakeholder groups. Full legend on page 1.

Item 7 - Study design

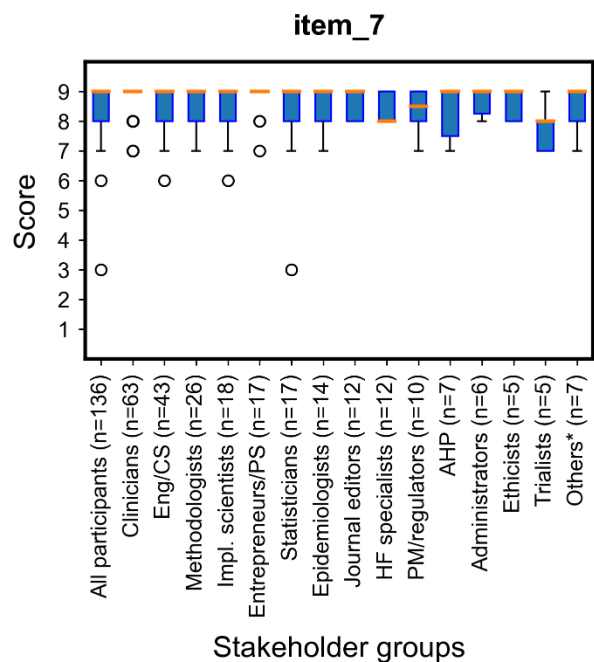
Describe the study design.

Overall median: 9

98.5 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 10

Summary of comments:

- Important item *but* possibly too brief/vague to be useful as an item
 - Overlap with methods items that follow
- Add detail to E&E document
 - Need standardised language
- Essential to establish confidence in the results.

Figure 9. Median score and IQR for item 7 of the revised list, by stakeholder groups. Full legend on page 1.

Item 8a - Participants

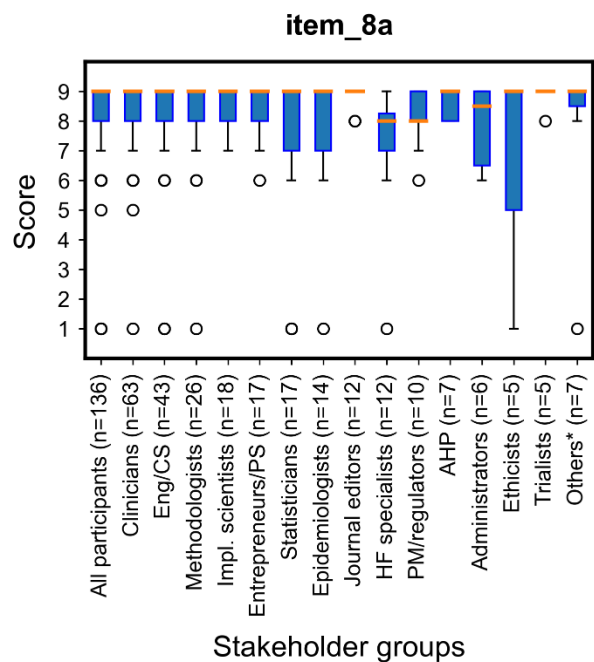
Describe precisely how patients were recruited, stating the inclusion and exclusion criteria, and how the number of recruited patients was selected.

Overall median: 9

93.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 17

Summary of comments:

- Is recruitment the right word if patient's data is from a database etc
- Is it the intended or actual numbers?
- Last sentence re: size is vague -> if referring to sample size then simply state this, or separate the item
- Intended number of patients is a statistical issue
- A flowchart/illustration would be useful
- Consider following changes:
 - Removing the word 'precisely'
 - Replacing 'selected' with 'chosen'.

Figure 10. Median score and IQR for item 8a of the revised list, by stakeholder groups. Full legend on page 1.

Item 8b - Participants

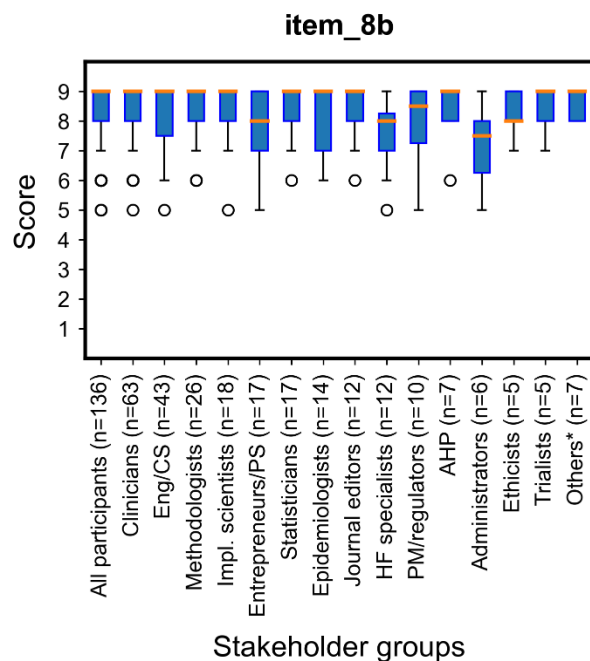
Describe precisely how users were recruited, stating the inclusion and exclusion criteria, and how the number of recruited users was selected. If applicable, describe the control group in sufficient detail to allow replication.

Overall median: 9

90.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 24

Summary of comments:

- Control group may not be necessary at the early stage, or move to 8a
- Rephrase to 'comparison group'
- 'Users' maybe too vague -> clinician, data collector etc?
- Is it the intended or actual numbers?
- Last sentence re: size is vague -> if referring to sample size then simply state this, or separate the item
- Intended number of patients is a statistical issue
- Consider:
 - Merging 8a and 8b
 - Add type and experience level of clinicians, demographics
 - Removing the word 'precisely'
 - Add mechanism of recruitment rather than just criteria
 - Replacing 'selected' with 'chosen'.

Item 8c - Participants

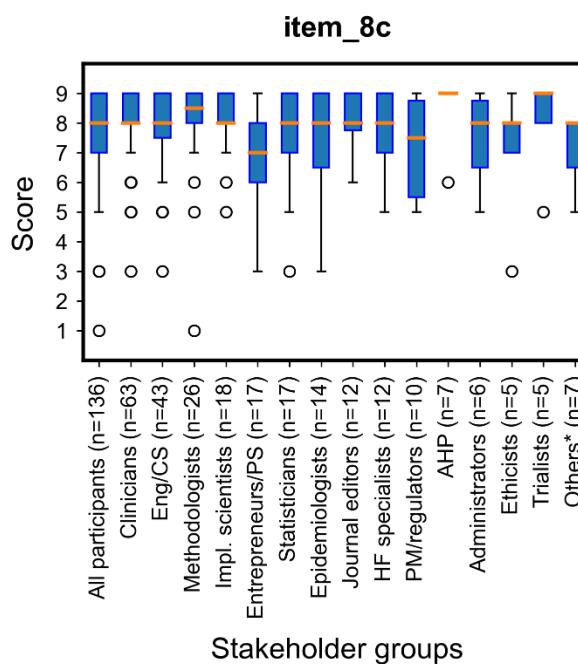
Describe any attempts to familiarise the users with the algorithm, including any training received.

Overall median: 8

85.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 18

Figure 12. Median score and IQR for item 8c of the revised list, by stakeholder groups. Full legend on page 1.

Summary of comments:

- Relevant but not essential / could cut / could move to human factors items
- Critically important as part of the description of the intervention
- Might be better as a separate line of research as requires more best practices and standards
- This is the sort of detailed data that is required in clinical evidence for regulatory approval, might be too much to ask here.
- Add if training is feasible in real-world as well as duration
 - Is the aim to produce capable users or medically generalisable/credible results?
- Phrasing clarity
 - One suggestion to add 'prior to use' after the word algorithm.

Item 9 - Algorithm

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.

Overall median: 8

89.6 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None

Number of comments: 30

Summary of comments:

- Just cite previous work / keep very brief
 - If no previous published work then may need to be a long section
- Version numbers may not be easily trackable, meaningful or available
- More important than the version number is if the algorithm has changed since the preclinical development studies
- Overlap with item 4
- Algorithm details and details about the training set population should not be mixed
- Brief description of the algorithm would be useful to identify these studies for systematic reviews
- Training data may not be on 'patients' per se
- Consider adding: dataset size, testing data, why were performance metrics chosen, case-mix
- Type of model used and details about the training set population are the two most important components here
- Rephrase second sentence: Refer to any previous work, including development and performance evaluation in validation/test data.

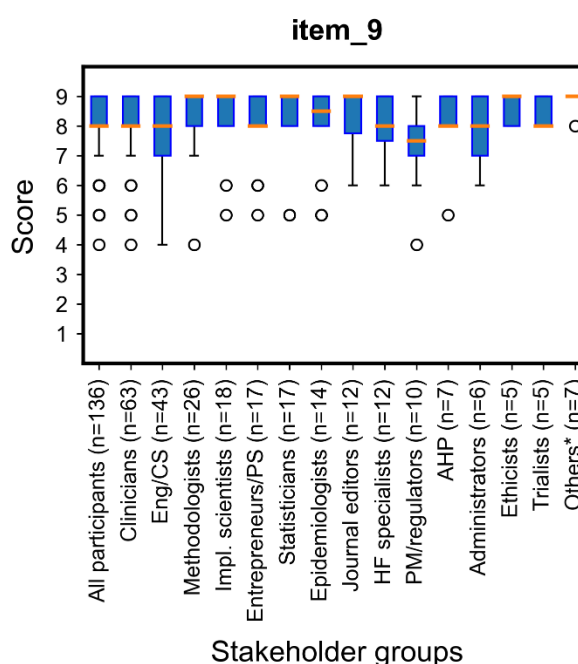


Figure 13. Median score and IQR for item 9 of the revised list, by stakeholder groups. Full legend on page 1.

Item 10a - Implementation

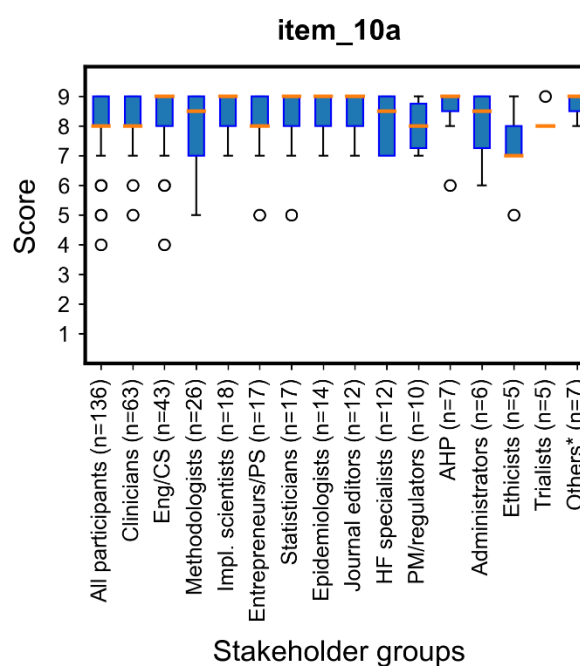
Describe precisely the environment in which the algorithm was tested, including the availability of the algorithm's input data and which additional clinical information (i.e. not provided by the algorithm) was accessible to the users to interpret or put into context the output of the algorithm.

Overall median: 8

94.8 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 18

Summary of comments:

- Consider removing "precisely" and replacing "environment" with "setting"
- Overlap with other items
- 'Additional info' not provided by algorithm might be too vague and difficult to collect - > e.g. EHR, gut feeling, guidelines
- Few comments stating that it is an important item (for example to assess difference between context)
- Availability of input data might be better suited in 10e (or add 'expected' prior to availability).

Figure 14. Median score and IQR for item 10a of the revised list, by stakeholder groups. Full legend on page 1.

Item 10b - Implementation

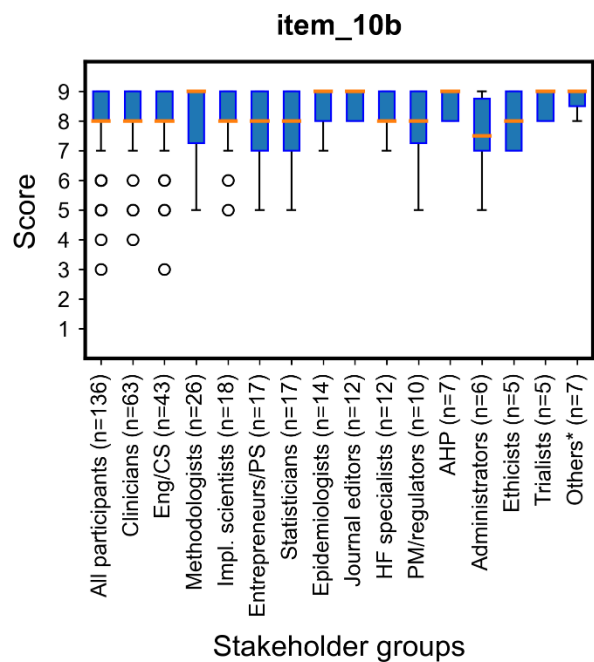
Describe the clinical workflow/pathway in which the algorithm was deployed and who held the responsibility for the final clinical decision.

Overall median: 8

91.0 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 26

Figure 15. Median score and IQR for item 10b of the revised list, by stakeholder groups. Full legend on page 1.

Summary of comments:

- Replace “who held the responsibility for the final clinical decision” by “how the final clinical decision was reached”, as this will cover shared decision making and conflict resolution
- Unclear if by decision it means who has authority and how determined, could be a group/committee rather than single person
 - What does ‘responsible’ mean -> legally? On paper?
- Consider:
 - Merge with item 10a/10c
 - Add flow-diagram
 - the term 'deployment' should be generally avoided as its military context strikes many clinical folks as unsavoury. 'Integrated,' 'tested,' 'implemented,' or 'utilized' work well.

Item 10c - Implementation

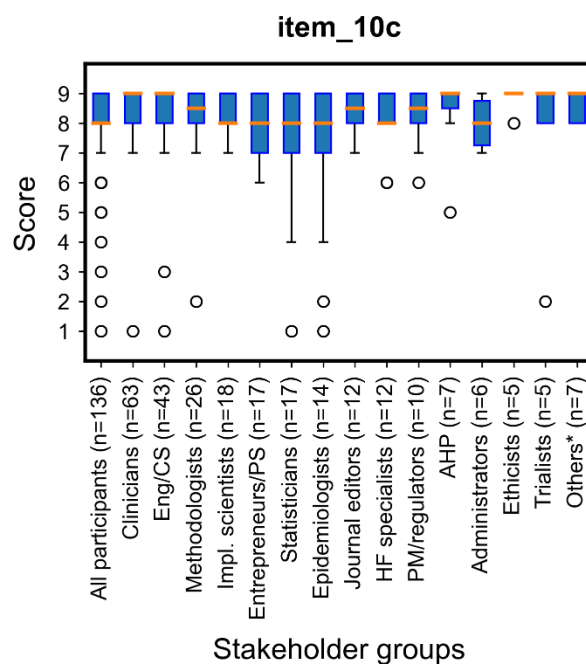
Describe precisely how the algorithm was used and the timing of the decision support.

Overall median: 8

93.4 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 25

Summary of comments:

- Could be merged with 10b (many comments)
- Is this called "concurrent reader paradigm"? "first reader paradigm" is confusing, the algorithm could be misunderstood as the main decision making while the user is only a "second opinion", simply reversing the roles of the "second reader paradigm".
- 'Timing' unclear -> could be interpreted by some as decision speed savings rather than where in the pathway i.e. one comment says not feasible practically
- Need to define 'decision support' as spectrum in autonomy between automated and support rather than binary
- Should make the difference between intended use (item 3) and actual use during the study (item 10c)
- Consider:
 - Remove 'precisely'
 - Separate how used from timing.

Figure 16. Median score and IQR for item 10c of the revised list, by stakeholder groups. Full legend on page 1.

Item 10d - Implementation

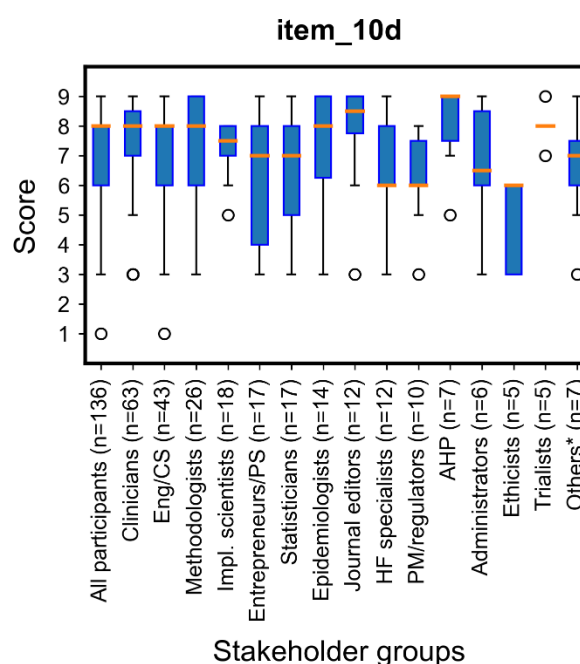
Describe the technical details of the implementation, including the integration within the existing study site IT infrastructure, the software and hardware needed to run the algorithm and any algorithmic thresholds used.

Overall median: 8

72.6 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- HF specialists
- PM/regulators
- Ethicists



Number of comments: 29

Figure 17. Median score and IQR for item 10d of the revised list, by stakeholder groups. Full legend on page 1.

Summary of comments:

- Thresholds probably belongs in a separate item (e.g. item 9 / 10b / 10c / 10e / 10f)
- Maybe too broad / too technical / relevance of hardware software questioned unless close to being distribution ready
- Early stage assessment may be on 'local' implementations rather than fully linked to ICT
- Differing opinions re: needed for replicability vs. most saying would take up too much space and better in supplement
- Concern that too much detail could drown out important technical aspects
- Some hospital infrastructure will be too diverse to generalize -> maybe higher level principles rather than minutiae e.g. feasibility of deployment.

Item 10e - Implementation

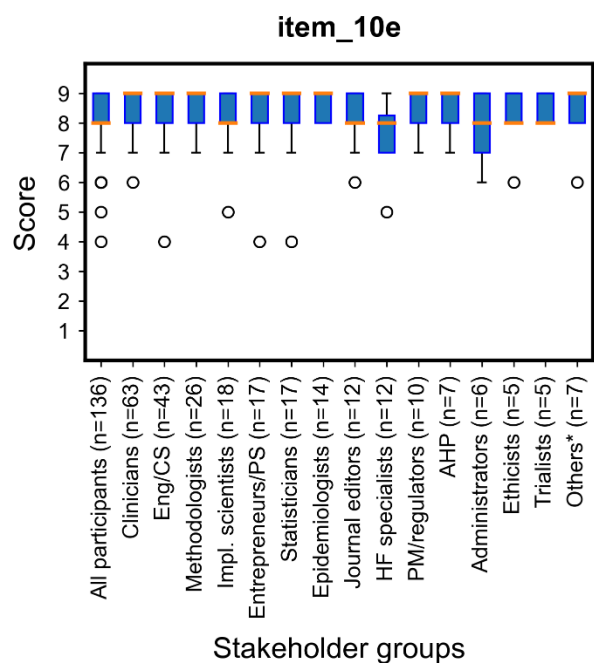
Identify the data used as inputs. Describe how the data were acquired, the process needed to enter the input data, any pre-processing applied and how missing/low-quality data were handled.

Overall median: 8

95.6 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 15

Summary of comments:

- Clarify the difference between data handling in development vs implementation
- Some overlap with 10a
- Some overlap potentially with development/validation study publication
- General feeling that reporting on missing data is important.

Figure 18. Median score and IQR for item 10e of the revised list, by stakeholder groups. Full legend on page 1.

Item 10f - Implementation

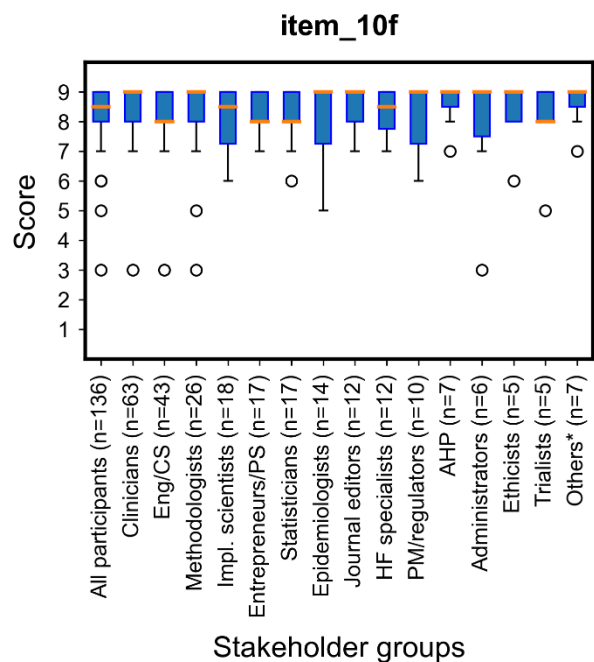
Describe the algorithm outputs and how they were presented to the users.

Overall median: 8.5

97.1 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 14

Summary of comments:

- Could request authors to include images
- Define 'users' more clearly
- Important as will also thereby report whether explainability demonstrated to users
- This is particularly important- especially in light of David Spiegelhalter's work on impact of risk presentation. The way risk predictions are presented can affect their decisions- dynamic/static information, and representations of variability or confidence in a point estimate
- Overlap with item 9, 10b, 11a.

Figure 19. Median score and IQR for item 10f of the revised list, by stakeholder groups. Full legend on page 1.

Item 11a - Outcomes

Specify the primary and secondary outcomes measured.

Overall median: 9

95.5 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None

Number of comments: 8

Summary of comments:

- Could merge with study design
- There may not be secondary outcomes.

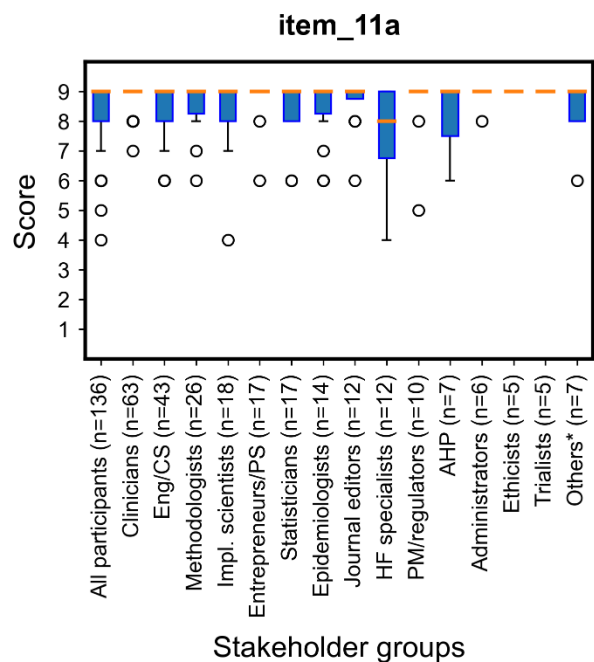


Figure 20. Median score and IQR for item 11a of the revised list, by stakeholder groups. Full legend on page 1.

Item 11b - Outcomes

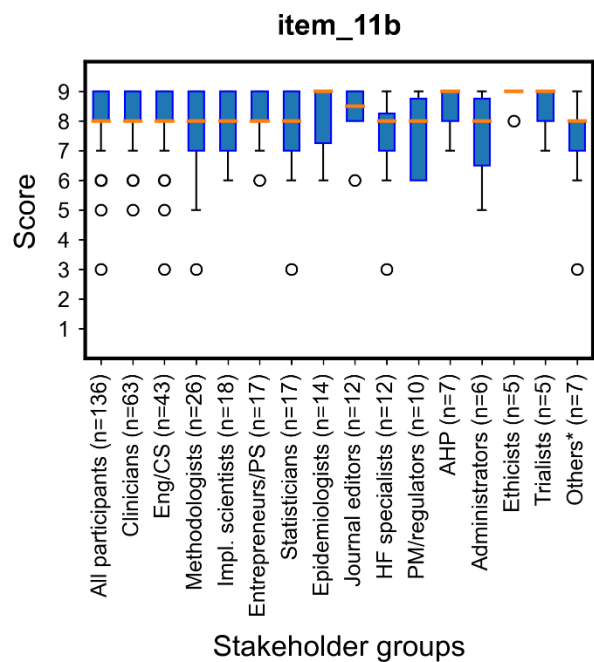
Describe how algorithm recommendation/output errors were defined and how they were identified.

Overall median: 8

92.6 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 13

Summary of comments:

- Very different from 11a, should not be under the same heading
- Wording unclear (does error relates to technical or clinical error?)
- Could merge with item 11a, or 13
- Consider splitting definition vs. identification
- Ground truth is often difficult to ascertain - reference standard can be imperfect so defining it is important
- Need rephrasing “recommendation/output errors” as it created some confusion amongst participants
- Error definition could be difficult/variable.

Figure 21. Median score and IQR for item 11b of the revised list, by stakeholder groups. Full legend on page 1.

Item 12 - Analysis

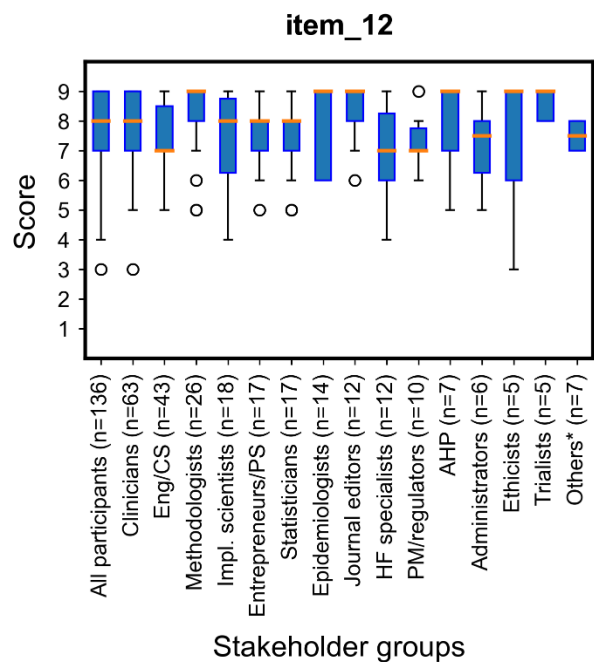
Describe the pre-specified analysis plan for the primary and secondary outcomes as well as for any prespecified additional analyses, including subgroup analyses and their rationale.

Overall median: 8

85.1 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 14

Summary of comments:

- Could merge with item 11a
- Too early to expect pre-specified analysis plan
- “prespecified” suggests a study protocol, which might not always be available
- Post-hoc analysis may also be useful
- May only be enough space for primary analysis
- Defining subgroup analysis is important.

Figure 22. Median score and IQR for item 12 of the revised list, by stakeholder groups. Full legend on page 1.

Item 13a - Safety

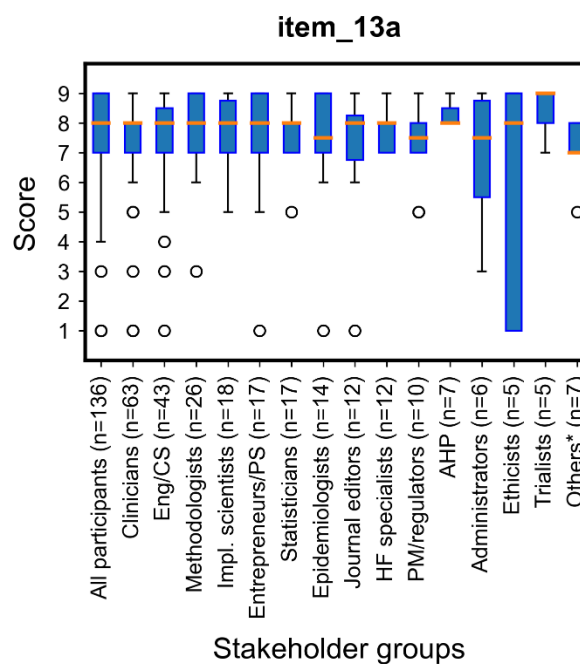
Define the algorithm safety requirements, how these were established preclinically, and how compliance to these requirements was evaluated during the study.

Overall median: 8

90.4 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 13

Summary of comments:

- Needs more definition/standardisation of terminology
- Could be subjective.
- Report whether PPI used to inform the process
- Reporting on risk management procedure is more important than reporting on safety requirements
- Could merge with item 13b
- Wording: how can safety be accurately “determined” pre clinically. “Modelled” maybe?
- Should be covered by ethics review.

Figure 23. Median score and IQR for item 13a of the revised list, by stakeholder groups. Full legend on page 1.

Item 13b - Safety

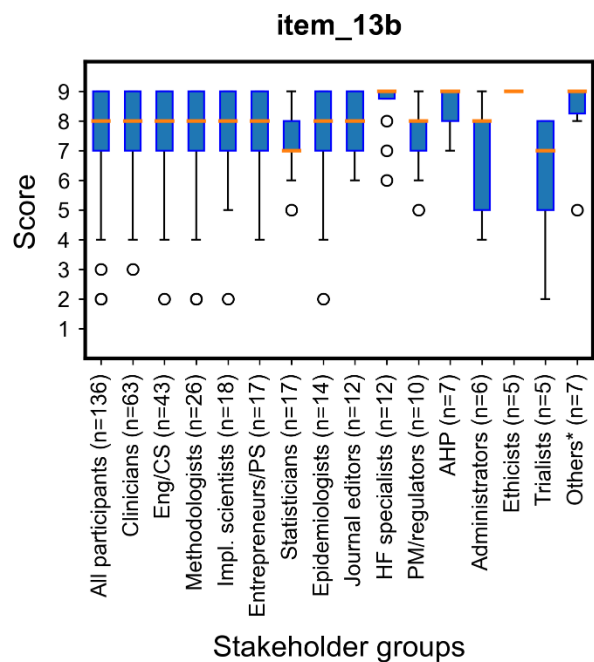
Describe the methodology used to detect any new, unexpected risks arising from the real-life clinical use of the algorithm.

Overall median: 8

84.2 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 20

Summary of comments:

- Is there really a methodology to detect new/unexpected risk? Consider rephrasing
- Vague / subjective
- Put in discussion instead of methods
- Mention reporting mechanism and mitigating actions too
- Could merge with item 13a
- Should have been covered in ethics approval
- Out of scope for early evaluation
- maybe limit to patient safety at early-stage.

Figure 24. Median score and IQR for item 13b of the revised list, by stakeholder groups. Full legend on page 1.

Item 14 - Human factors

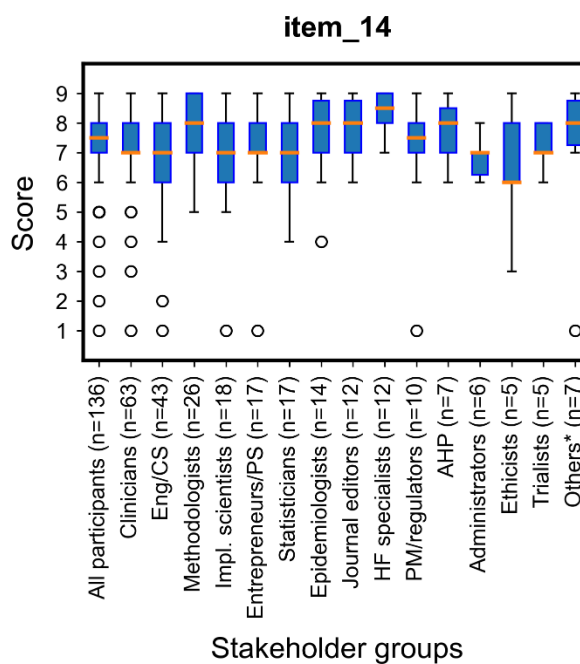
Describe the human factors tools, methods or frameworks used, the use cases considered and the users involved in the human factors evaluation.

Overall median: 7.5

76.2 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 23

Figure 25. Median score and IQR for item 14 of the revised list, by stakeholder groups. Full legend on page 1.

Summary of comments:

- Merge with outcomes (items 11a/b), merge with item 12
- Too general/vague/abstract, overlaps with other sections (e.g. workflow/pathway)
- Perhaps better in supplement, may take too much space
- HF work is specialised, complicated and expensive. It shouldn't be a barrier to assessing whether there is clinical utility which will then be more formally assessed.
- Definitions for various terms in the item is required, add more detail to E&E document
- Standard HF terminology would be important
- Add 'if applicable'
- Should this be separate study / is it essential for early-stage?

Item 15 - Patient engagement

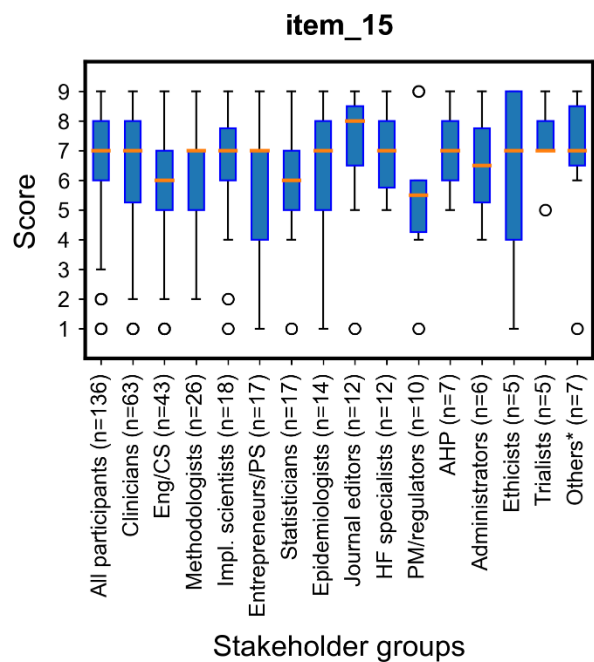
State whether patients were involved in any aspect of the study design, conduct or in the development of the research question or outcome measures.

Overall median: 7

57.1 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 28

Figure 26. Median score and IQR for item 15 of the revised list, by stakeholder groups. Full legend on page 1.

Summary of comments:

- Useful but not essential for early stage evaluation
- Could be rephrase/merged with 16 as a stakeholder involvement item (patients, public, ethicists, end users, ...)
- Important as adds nudge to consult patients/change the paradigm toward more consideration for the end beneficiaries
- Already required by many funding bodies and journals
- Overall: big split between support and those who feel too early/not necessary.

Item 16 - Ethics consideration

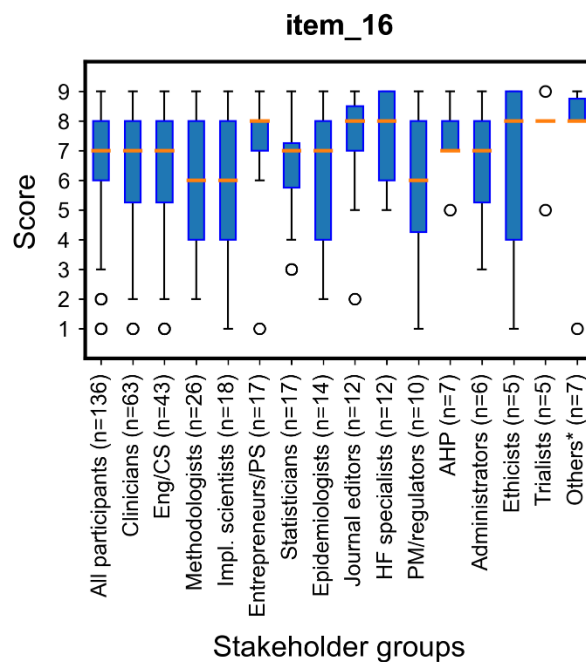
Describe any ethics methodology, consultation or involvement during the design or implementation of the study.

Overall median: 7

61.7 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 26

Summary of comments:

- Merge with item 6a / already covered by ethics there
- Redundant as covered by ethics/IRB processes and if the main focus of an actual study then reasonable to describe in methods, else add to discussion section
- Too broad / vague -> consider rephrase: "Describe how ethical issues were identified and addressed during the design and implementation of the study"
- These could include algorithmic fairness assessment, consultation with vulnerable/marginalized groups, rationale for selection of methodology with respect to equity or other fairness concepts, whether IRB/REBs approved waivers of consent for the research participants, etc. A few other items or considerations could be captured under this item if we need to reduce the item count.
- Could be combined with item 15 under a general stakeholder involvement item.

Figure 27. Median score and IQR for item 16 of the revised list, by stakeholder groups. Full legend on page 1.

RESULTS

Item 17a - Participants

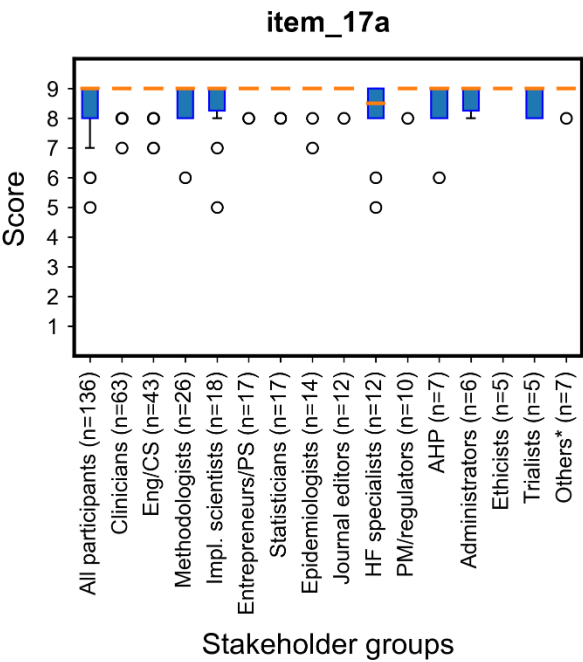
Describe the patient study group baseline characteristics (number, number of centres, age, sex, ethnicity if relevant, comorbidities, prevalence of the target conditions, etc.).

Overall median: 9

97.8 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 5

Summary of comments:

- Avoid 'etc' and provide short list of required criteria.

Figure 28. Median score and IQR for item 17a of the revised list, by stakeholder groups. Full legend on page 1.

Item 17b - Participants

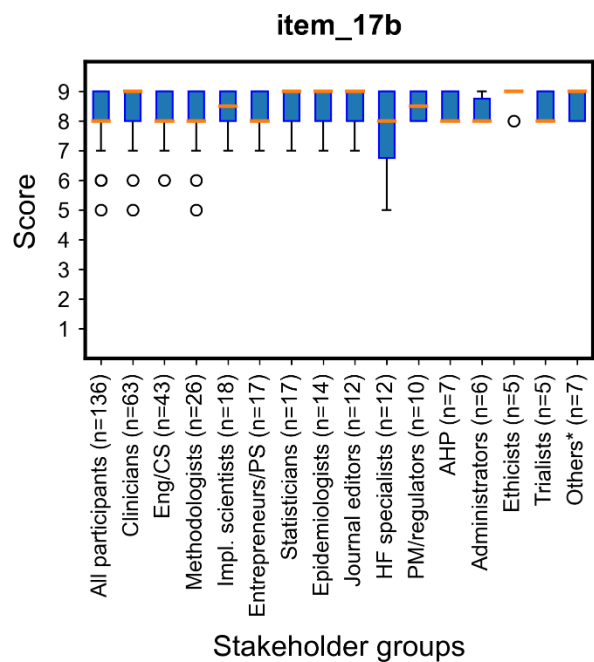
Describe the users study group baseline characteristics (number, number of centres, specialty, seniority, previous experience with digital support, etc.).

Overall median: 8

96.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 6

Summary of comments:

- Avoid being too prescriptive with wording (e.g. previous experience may be difficult to quantify)
- Avoid 'etc' and provide short list of required criteria.

Figure 29. Median score and IQR for item 17b of the revised list, by stakeholder groups. Full legend on page 1.

Item 18a - Implementation

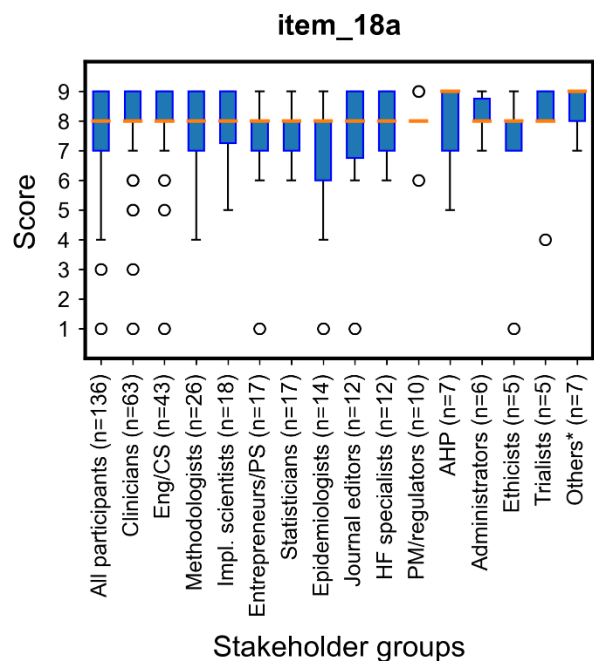
Report on the user exposure to the algorithm (implementation reach), on the number of instances the algorithm was used (implementation dose) and on the users' adherence to the intended implementation (implementation fidelity).

Overall median: 8

86.7 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 8

Summary of comments:

- Terms (e.g. fidelity) need to be defined more clearly (avoid specialty/expert jargon)
- If adherence is a mandatory outcome, it should be defined in the methods section
- Probably the most important data point in the early clinical phase.

Figure 30. Median score and IQR for item 18a of the revised list, by stakeholder groups. Full legend on page 1.

Item 18b - Implementation

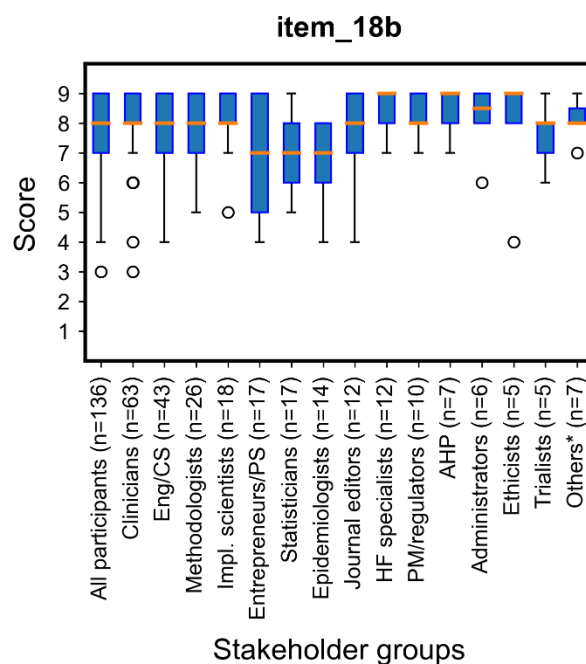
Report changes caused by the algorithm to the clinical workflow, if any.

Overall median: 8

85.9 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 13

Summary of comments:

- Dependant on how implemented / difficult to quantify
- Should report as an outcome (item 20a) or if unintended then as an unintended risk (21b)
- most clinical safety, ethical issues, etc. will derive from shifts in the clinical pathway rather than the design of the algorithm itself
- Unclear -> does it mean changes required to implement algorithm or changes arising from algorithm outputs/downstream effects
- algorithm performance and workflow change may not be assessed in the same study
- Unclear why the language here has gone from "describe" to "report" - describe is fine
- Major deviations from clinical workflow will be important to note and report on but not every small deviation is important.

Figure 31. Median score and IQR for item 18b of the revised list, by stakeholder groups. Full legend on page 1.

Item 19 - Modifications

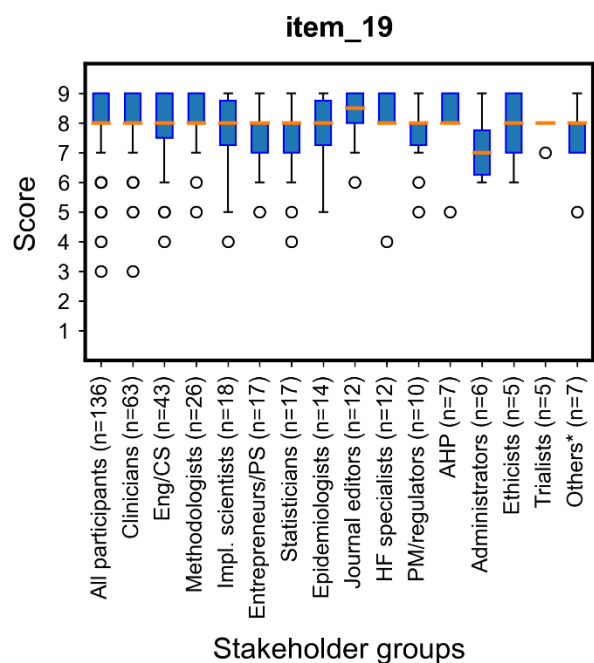
Report any changes made to the algorithm or its hardware platform between the prototype used at the beginning of the study and its final version. Report the timing of these modifications and the changes in outcomes observed after each of them.

Overall median: 8

88.9 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 15

Summary of comments:

- Add 'if applicable'
- Also explain the reasons for changes "Report the timing of these modifications, [reasons for modification], and the changes in outcomes observed after each of them."
- Clarify how this would impact continuously learning algorithms
- The software and hardware may undergo any number of changes and updates. Requiring outcome analyses after every update is not reasonable.
- Definitions could be clearer (i.e. prototype).

Figure 32. Median score and IQR for item 19 of the revised list, by stakeholder groups. Full legend on page 1.

Item 20a - Main results

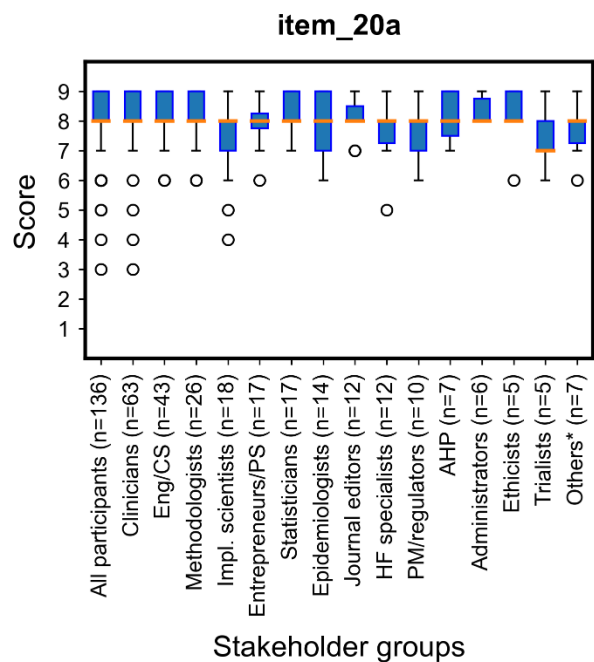
Report on the prespecified outcomes for the algorithm-assisted users (both overall and at an individual user level), including any variation over time.

Overall median: 8

92.2 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 12

Figure 33. Median score and IQR for item 20a of the revised list, by stakeholder groups. Full legend on page 1.

Summary of comments:

- Could rephrase 20a and b for clarity. For example 'Report on the prespecified outcomes. This should be both for algorithm-assisted users (i.e. human/AI combination) and, where applicable, for the stand-alone algorithm. For algorithm-assisted users this should be both overall and at individual user level, with any variation over time recorded.'
- Definitions could be tighter
 - Several comments on this item being unclear
- Individual user level may be too much detail -> maybe just by pre-specified subgroup
- 'Variation in time' only if actually planned as an analysis.

Item 20b - Main results

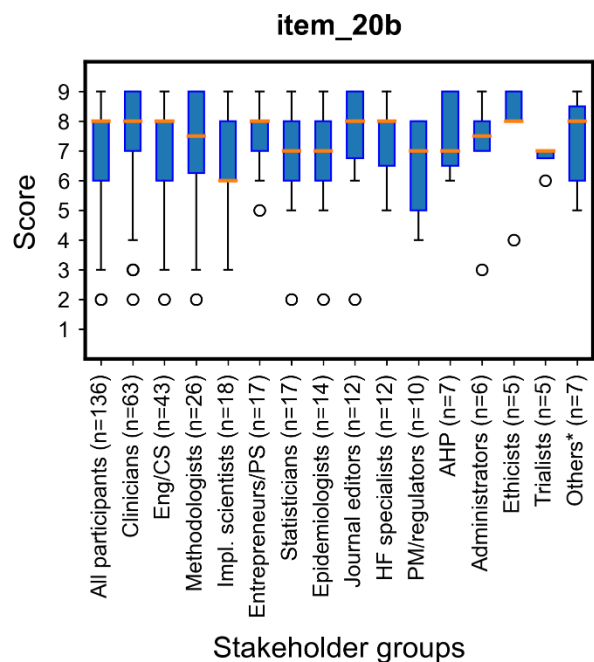
Report on the prespecified outcomes for the stand-alone algorithm, if applicable.

Overall median: 8

74.0 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- Implementation specialists



Number of comments: 18

Summary of comments:

- Merge with item 20a +/- 20c
- Hypothetical performance is a slippery slope and should not be the focus of live early clinical evaluation
- Already covered by previous in-silico work
- What is most important is real impact that also includes implementation and uptake
- Wording not clear
- Could combine with item on human errors.

Figure 34. Median score and IQR for item 20b of the revised list, by stakeholder groups. Full legend on page 1.

Item 20c - Main results

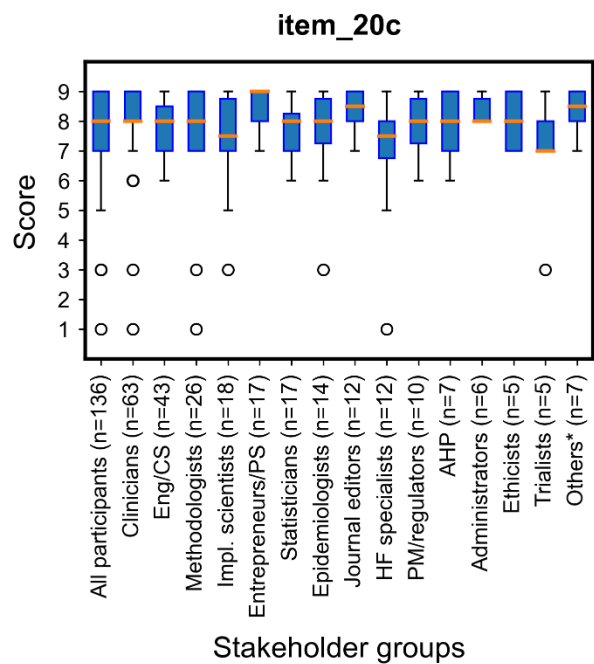
Report on the prespecified outcomes for the control group, if applicable.

Overall median: 8

91.0 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 10

Summary of comments:

- Merge with item 20a +/- 20b
- Control group not necessary at this stage of evaluation
- Control group necessary to interpret the added value of the algorithm.

Figure 35. Median score and IQR for item 20c of the revised list, by stakeholder groups. Full legend on page 1.

Item 21a - Safety and errors

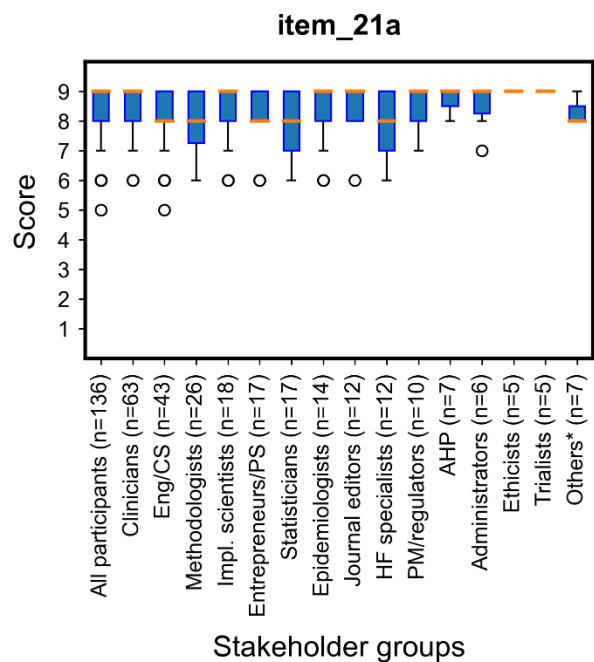
Report on the compliance with the specified safety requirements and any severe adverse events.

Overall median: 9

93.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 4

Summary of comments:

- Merge with item 21b
- Separate out compliance and severe adverse events
- Consider all adverse events rather than just severe.

Figure 36. Median score and IQR for item 21a of the revised list, by stakeholder groups. Full legend on page 1.

Item 21b - Safety and errors

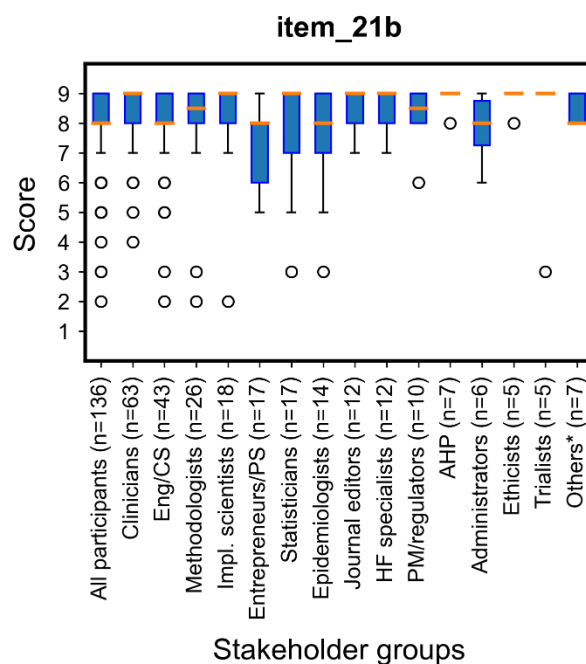
Report any additional risks identified from the real-life clinical use of the algorithm.

Overall median: 8

91.1 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 17

Summary of comments:

- Merge with item 21a
- Important but difficult to ensure, may be difficult for results section as opposed to discussion
- Also report type and quantity
- This could become a long list of potential risks. I think what is very important is to state its limitations based on the design or results of the study. For example, the algorithm works in only specific patient types (e.g. early stage cancer but is unreliable in late stage)
- Consider opposite too -> unexpected real-world benefits
- Only additional risks which can be credibly described as possibly associated with the use of the algorithm.

Figure 37. Median score and IQR for item 21b of the revised list, by stakeholder groups. Full legend on page 1.

Item 21c - Safety and errors

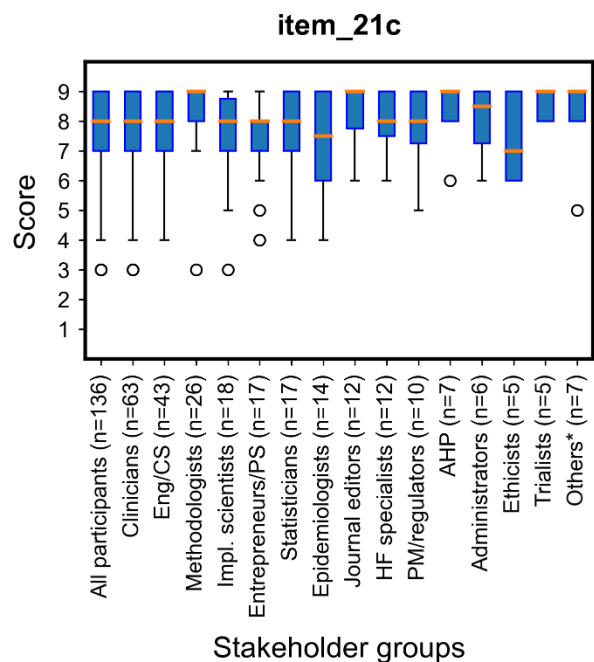
Report any algorithm malfunction or issues with hardware or software during the study.

Overall median: 8

83.0 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 16

Summary of comments:

- Merge as subset of item 21b
- Move to discussion
- More of an item for regulatory approval
- Should only be reported if having an impact on the outcomes
- Report consequences too as well as how users overcame issues
- How much granularity is required? Not reasonable to go through all patches/updates/bugs.

Figure 38. Median score and IQR for item 21c of the revised list, by stakeholder groups. Full legend on page 1.

Item 21d - Safety and errors

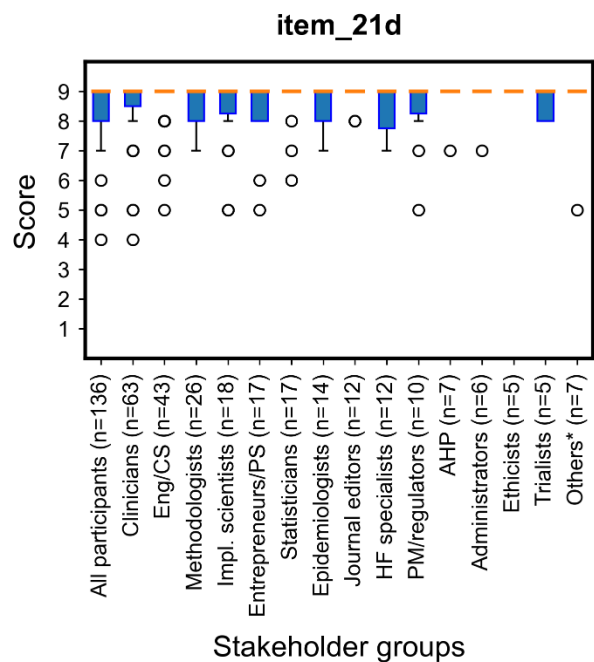
Report any algorithm recommendation errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual impact on patient care.

Overall median: 9

97.0 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 10

Summary of comments:

- Ideally in a standardised format
- Overlap with item 21a/21b and other safety points
- Too difficult to specify exactly what to document here -> beyond manuscript scope
- A grading of errors is also necessary (e.g. critical vs. non).

Figure 39. Median score and IQR for item 21d of the revised list, by stakeholder groups. Full legend on page 1.

Item 21e - Safety and errors

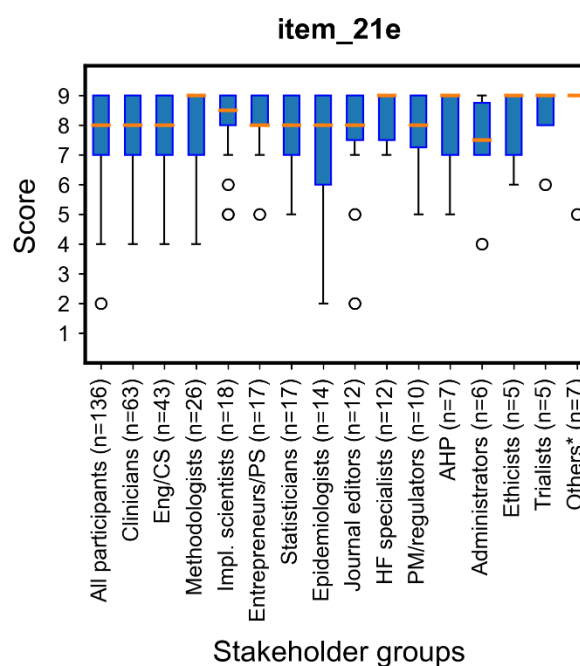
Report any human errors, detailing their rate of occurrence, causes, whether they were corrected and potential/actual implication for patient care.

Overall median: 8

84.8 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 22

Summary of comments:

Figure 40. Median score and IQR for item 21e of the revised list, by stakeholder groups. Full legend on page 1.

- very difficult to obtain and quantify; many errors will be undetected unless there is an oversight board looking into all decisions made
- Human errors would need careful prior definition, who would define them? As humans are often the gold standard, it can be difficult to assess their errors. No clear and accepted definition so far
- may also need ethics/IRB so users know this data is being captured
- is this human error in using the system (protocol violation, data input)? Or human error in clinical judgement whether using the system. Decision errors much harder to measure.
- human and algorithm errors might not always be differentiated => decision errors
- Perhaps combine 21d, 21e or 23a? Report on any errors, detailing whether they are algorithm recommendation errors, human errors, rate of occurrence... etc.
- As for performance, the initial evaluation of a technology which identifies a challenge around human errors should not be definitive in this type of study
- Consider the term "use error" rather than "human error"
- Physicians' errors in their routine clinical work is not a part of the algorithm evaluation. This question is either confusing or redundant.
- Overlap with reporting of control group outcome.

Item 22 - Subgroup analysis

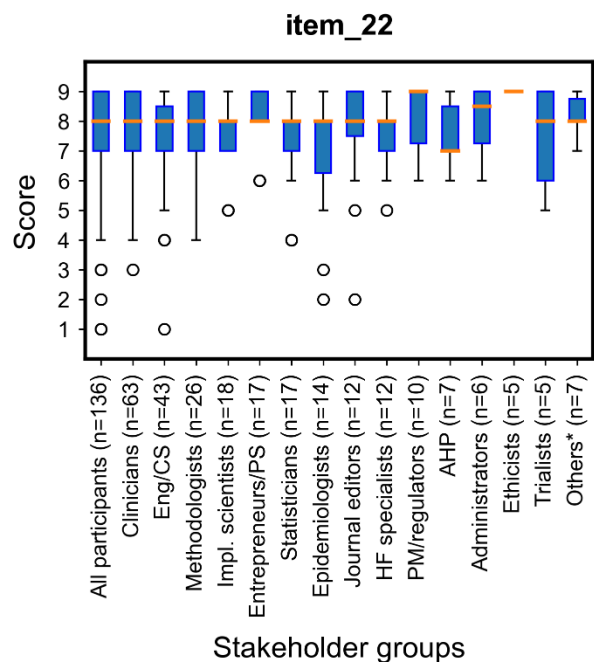
Report on the difference in the main outcomes according to the specified subgroups.

Overall median: 8

77.9 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 24

Summary of comments:

- Only if sample size appropriate as per analysis plan and multiple hypothesis testing etc taken into account
- Risky to encourage mandatory subgroup analyses, maybe less relevant for early-stage
- Add 'if applicable'
- Merge with main outcome items and add 'including any specified subgroups'
- Could be captured under item 16 Ethics Considerations. An important consideration here would be to link any pre-described algorithmic fairness methods to the observed prediction pattern in the prospective trial.

Figure 41. Median score and IQR for item 22 of the revised list, by stakeholder groups. Full legend on page 1.

Item 23a - Human factors

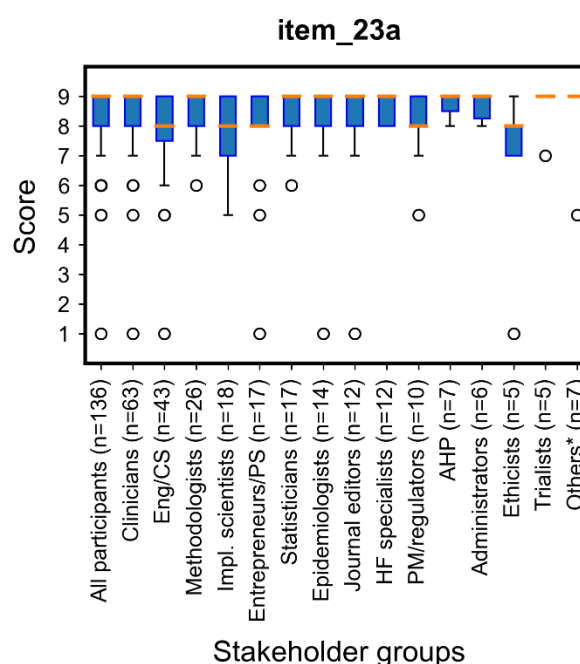
Report on the user agreement with the algorithm. Describe any instances of and reasons for user deviation from the algorithm's recommendations and, if applicable, user changing their mind based on the algorithm recommendations.

Overall median: 9

92.5 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 17

Summary of comments:

- May not be easy to collect/logistics -> asking users about it will affect the results so dependent on how study is designed
- Sometimes even impossible (for example in cases of combined output) => add "if applicable"
- why not asking the other way round - report on algorithm agreement with the user
- should ideally be linked to the reporting on algorithmic errors and final outcomes (Overlap with item 20b and 21d)
- Is this an outcome? If so define in the appropriate item (11) and report on it, else leave out (or move to introduction/discussion).

Figure 42. Median score and IQR for item 23a of the revised list, by stakeholder groups. Full legend on page 1.

Item 23b - Human factors

Report on the evolution of users' trust in the algorithm.

Overall median: 7

55.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- Entrepreneurs/private sector
- Journal editors
- Ethicists
- Trialists

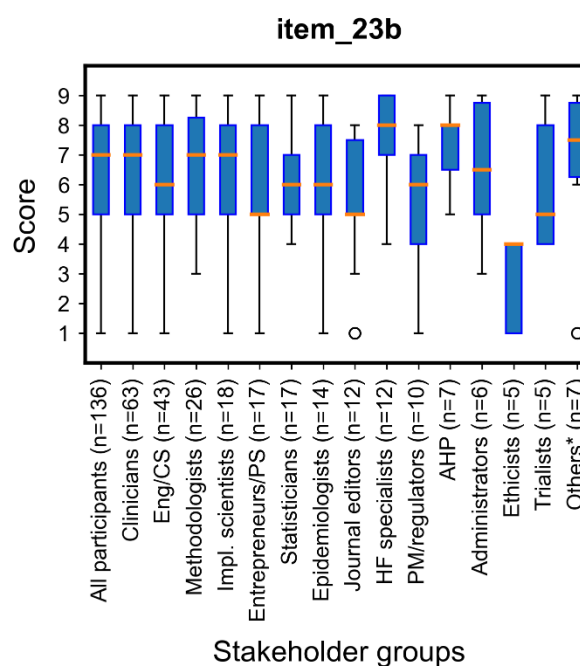


Figure 43. Median score and IQR for item 23b of the revised list, by stakeholder groups. Full legend on page 1.

Number of comments: 26

Summary of comments:

- Overlap with item 23a
- Can it be teased apart from familiarity over time?
- Subjective concept and no broadly accepted methods to measure trust so far
- As trust is rarely measured little is known about how to best measure it in trial
- Measuring evolution implies measuring the level of trust at every time point, which might not be feasible
- Study might be underpowered to measure evolution in trust reliably
- Depends on many potential unmeasured confounders
- Separate line of research, maybe not early-stage
- Merge human factors items together
- One of the most important issues to be resolved by studies in this phase of research.

Item 23c - Human factors

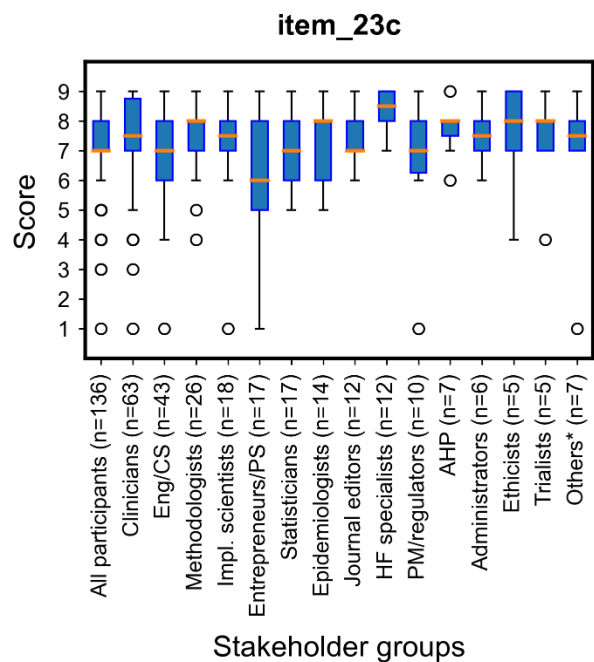
Report on the usability evaluation, including time to task completion, display interface evaluation and user satisfaction.

Overall median: 7

77.0 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 18

Summary of comments:

- Too detailed (time to task completion may not be relevant for any given task)
- These usability assessments may come at a later stage after proof of principle re: efficacy
- Too difficult for small scale early stage studies
- Separate line of research, out of scope
 - Impact evaluation and usability assessment probably in different papers
- Merge human factors items together
- Important to inform need for improvement in human computer interface.

Figure 44. Median score and IQR for item 23c of the revised list, by stakeholder groups. Full legend on page 1.

Item 23d - Human factors

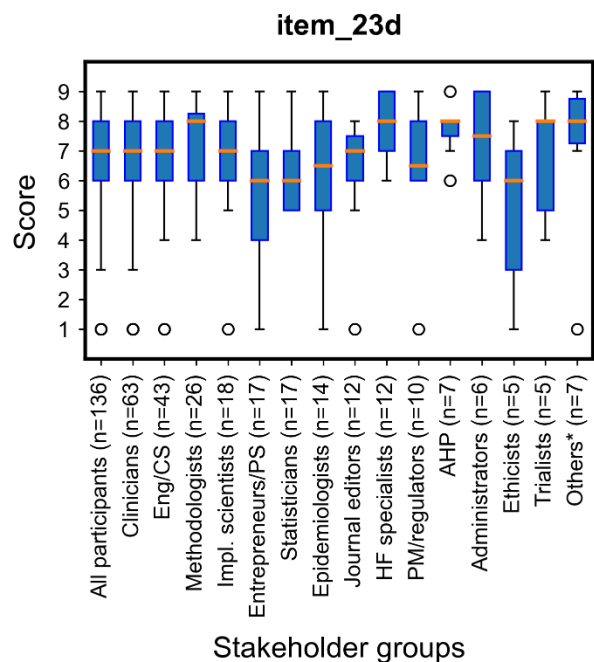
Report on the user workload and learning curves evaluation.

Overall median: 7

61.9 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 20

Summary of comments:

- There are plenty of examples of the impact of learning curves on the effect of technology.
- Ease of use is important to get as many people using it to get the technology 'rolled out'
- Hard to evaluate in early phase
- Does the item imply a change in workload over time? If yes the study might be underpowered to measure it.
- Very important but may be difficult to disentangle from trust curve
- May be too much to ask. This could be a different study, a reference would be enough.

Figure 45. Median score and IQR for item 23d of the revised list, by stakeholder groups. Full legend on page 1.

Item 23e - Human factors

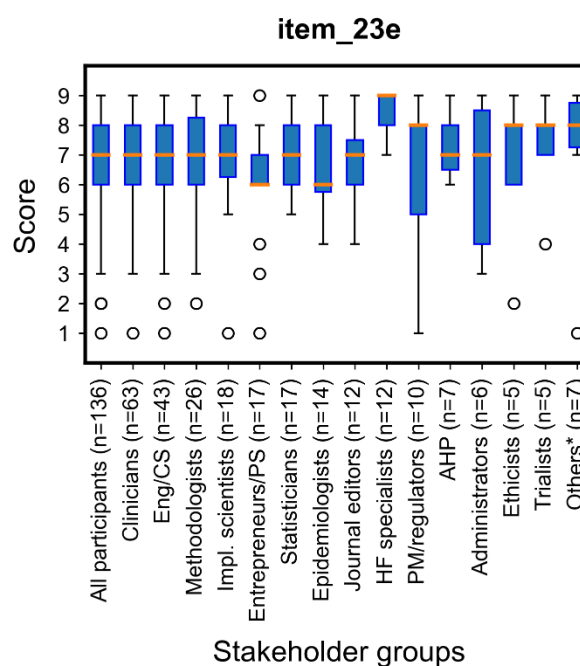
Report on the user perception of the algorithm outputs' interpretability and clinical value.

Overall median: 7

67.1 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 24

Summary of comments:

- Results could become very generic or subjective if method on how to collect this data is not specified (e.g. Likert scale).
- Might be too much for one paper -> depends on main aim of study
 - Separate line of research
 - A reference would be enough
- Overlap with item 23a, 23c, 23d (merge human factors items together)
- Some comments that interpretability not as important as efficacy and that we should be guided by the evidence of model efficacy, not how valuable users thought it was.
- Better in discussion section
- Separate out interpretability from clinical value. Perception of clinical value is more important than interpretability
- all of above for me as a clinician are key. This needs to make the patient and my life, better and easier.

Figure 46. Median score and IQR for item 23e of the revised list, by stakeholder groups. Full legend on page 1.

DISCUSSION

Item 24 - Support intended purpose

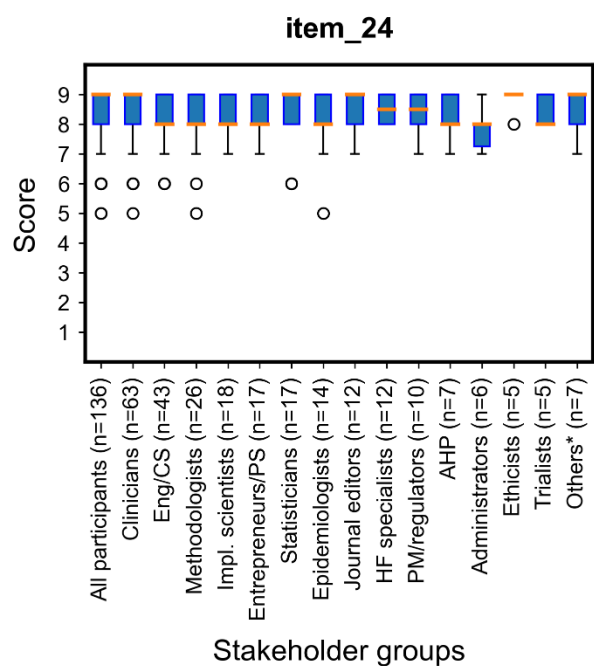
Discuss whether the obtained results support the intended purpose of the algorithm in real world clinical settings.

Overall median: 9

95.6 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 8

Summary of comments:

- should read "the results obtained in real world clinical settings support ..."
- Definitive pronouncement is not the objective of the study
- Not crucial, the readers can/should decide for themselves how to interpret the study
- may call this heading 'implications for practice'.

Figure 47. Median score and IQR for item 24 of the revised list, by stakeholder groups. Full legend on page 1.

Item 25 - Safety and errors

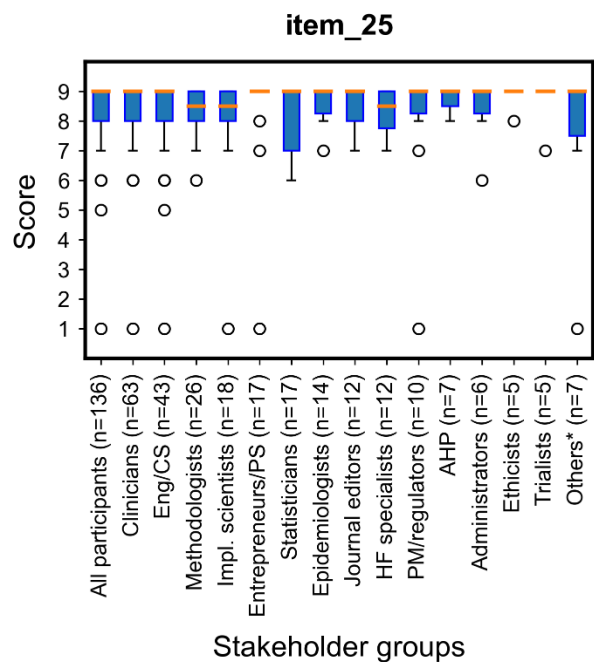
Discuss what the results suggest about the safety profile of the algorithm. Discuss the algorithm's errors and, if appropriate, identify any underlying pattern or algorithmic bias, explain how these can be mitigated..

Overall median: 9

97.0 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 9

Summary of comments:

- Could come under the item 16 (ethics)
- The focus should be on errors rather than safety
- Beyond the scope of initial clinical evaluations. This is a separate line of research.
- it's sufficient for us to ask authors to summarize their key findings, taking into account efficacy, human factors, and safety. (Could merge with other discussion items as: 'summarize their key findings, taking into account efficacy, human factors, and safety')
- Not crucial- it is subjective opinion of the investigators. Reader can also decide if study adequately described.
- Should the question be specifically about bias, as understanding systematic bias important
- this may give new and usable info and is different in this sense from item 24.

Figure 48. Median score and IQR for item 25 of the revised list, by stakeholder groups. Full legend on page 1.

Item 26 - Human factors

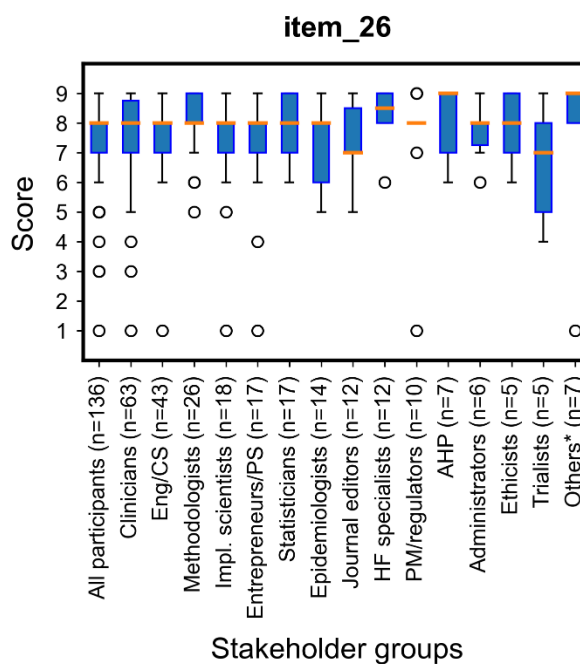
Discuss the results of the human factors evaluation and the reasons for human deviation from the algorithm's recommendations or intended use.

Overall median: 8

84.4 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 12

Figure 49. Median score and IQR for item 26 of the revised list, by stakeholder groups. Full legend on page 1.

Summary of comments:

- 'human deviation' suggests that the algorithm's output is the de facto decision that should be followed, which might not always be the case.
- "deviation might" be seen to a 'loaded' word, maybe "non concurrence".
- Should be reported only if the human factors are part of the study and there is enough information or results to support possible conclusions.
- Beyond the scope of initial clinical evaluations. This is a separate line of research.
- it's sufficient for us to ask authors to summarize their key findings, taking into account efficacy, human factors, and safety.
- Not crucial- it is subjective opinion of the investigators. Reader can also decide if study adequately described.
- Should be consistent with results section (and ideally aggregation of the results items, as dictating what should be reported in terms of outcomes is outside the scope of a checklist).

Item 27 - Scale up

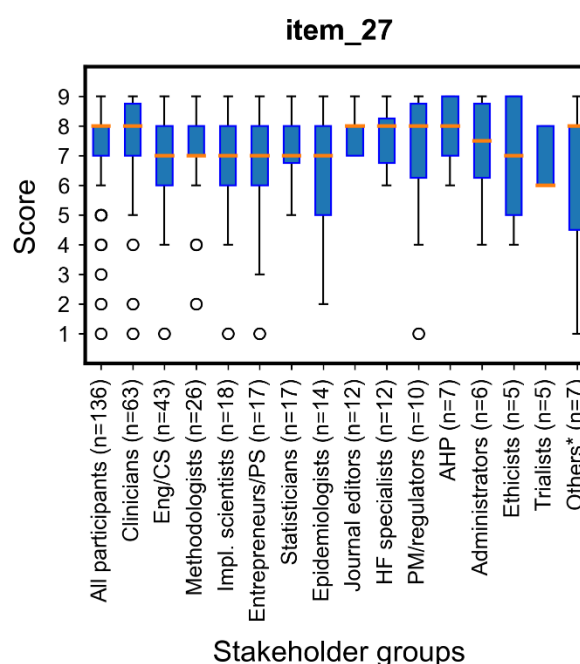
Discuss the scale-up feasibility and requirements, as well as the possible design of large-scale summative evaluation in light of the obtained results. Summarise the lessons learned from the study.

Overall median: 8

78.4 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- Trialists



Number of comments: 13

Summary of comments:

Figure 50. Median score and IQR for item 27 of the revised list, by stakeholder groups. Full legend on page 1.

- Range from: this is the main main purpose of study to discussion about future studies being out of scope
- Should we include something along the lines of generalisability. Scale-up is only one of the potential relevant dimensions to explore (could also be for example deployment in another clinical environment)
- "summarise the lessons learned from the study" sounds much more general than the previous sentence and does not add much to the other items already included in the discussion section
- More detail on how summative is defined in this context (e.g. might be small for some users studies for medical device)
- This will be highly speculative
- Lessons learned most important.
- not essential for the interpretation of the study
- Important to view scale-up as a separate study phase using the learning from early clinical studies to determine key design aspects.

Item 28 - Strength and limitations

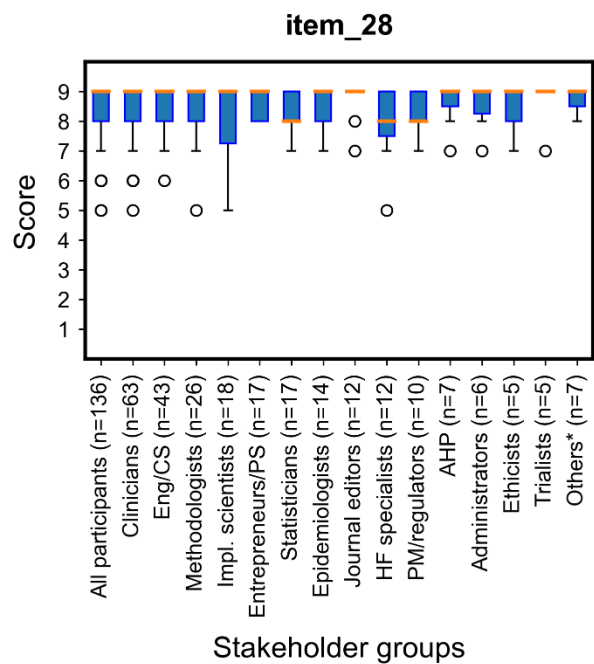
Discuss the strengths and limitations of the study, including any bias in the study design.

Overall median: 9

96.2 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 5

Summary of comments:

- Differentiate between training/data bias and study bias
- An obvious statement so why include ?

Figure 51. Median score and IQR for item 28 of the revised list, by stakeholder groups. Full legend on page 1.

STATEMENTS

Item 29 - Conflicts of interest

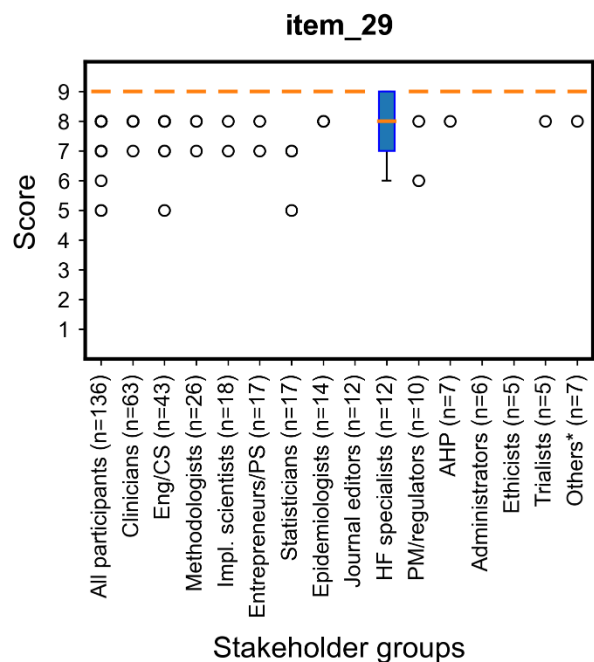
Disclose any relevant conflict of interest, including: the source of funding for the study, the role of funders, any other role played by commercial companies and authors' conflicts of interest.

Overall median: 9

97.8 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 7

Summary of comments:

- Treat commercial entities and government agencies as potential COIs
- An obvious statement so why include?
- Funder role is an important consideration.

Figure 52. Median score and IQR for item 29 of the revised list, by stakeholder groups. Full legend on page 1.

Item 30 - Data Availability

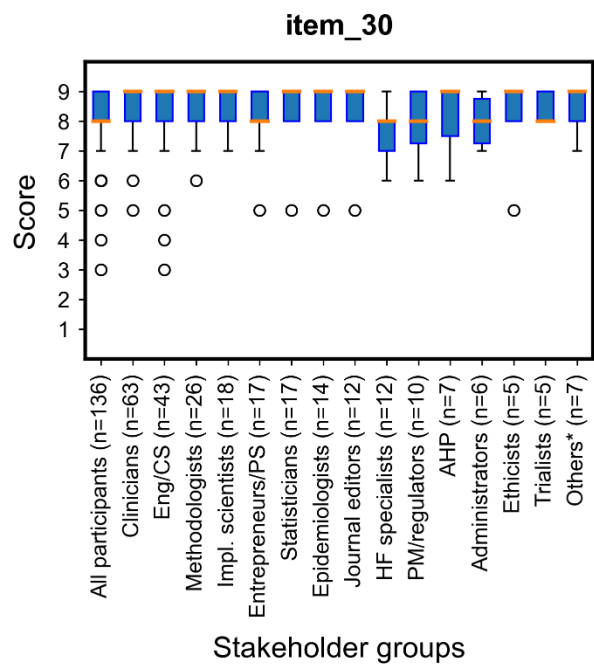
Disclose if and how data and code (pre-processing and algorithm) are available.

Overall median: 8

93.3 % of participants scoring ≥ 7

Stakeholder group whose median differs ≥ 2 points from overall median:

- None



Number of comments: 8

Summary of comments:

- This is a critical component of transparency and reproducibility.
- Include the license for which any data or code is available
- It's okay if not available but discuss it
- this should not be specified by guidelines - those who want to find out will do so
- Some individual data may not be releasable.

Figure 53. Median score and IQR for item 30 of the revised list, by stakeholder groups. Full legend on page 1.

GENERAL COMMENTS

Number of comments: 27

Summary of comments:

- Some items too generic and obvious, too many items could dilute important ones, could become unwieldy
- Subitem division not always logical, double barrelled items not always clearly related
- It would be useful to compare with other AI reporting guidelines in a table
- Wording tends to suggest algorithm supremacy, which might not always be the case.
- Generic stage approach may be too broad, IDEAL comparison not as straightforward as that is narrower in scope
- Checklist vs. how-to guide: should balance rigour with practicality
 - Consider short checklist, long E&E document and separate how-to guide
- Consider differentiating between « nice to have » and « minimum standard »
- Scope still not totally clear
- Could add a new 10g item on way risks are presented and whether clinical suggestion included
- Need a glossary of terms used
- May be difficult to report all of this in journal word count
- List much improved vs Round 1
- Seems to capture all the relevant aspects.