

# Consensus Meeting Minutes

14.06.21 – 16.06.21

M. Nagendran, B. Vasey

## DAY 1

---

- **BC opened and introduced the session**
- **Present:**
  - **BC** - Bruce Campbell - **(Chair, non-voting)**
  - **GC** - Gary Collins
  - **SD** - Spiros Denaxas
  - **BG** - Bart Geerts
  - **XL** - Xiao Liu
  - **BM** - Bilal Mateen
  - **MM** - Melissa McCradden
  - **PMc** - Peter McCulloch
  - **LM** - Lauren Morgan
  - **JO** - Johan Ordish
  - **CR** - Campbell Rogers
  - **SS** - Suchi Saria
  - **DT** - Daniel Ting
  - **BV** - Baptiste Vasey
  - **WW** - Wim Weber
  - **PW** - Peter Wheatstone
- **Apologies:**
  - **PM** - Piyush Mathur
  - **CR** - Campbell Rogers (from 15:00 to 16:00 BST)
- **Introduction of participants to each other**
- **BC:** Decisions made at this stage not absolutely final - may need to return to some decisions based on others where interconnected. Glossary of terms might be useful, **WW** as a journal editor agreed.
- **Vote held** -> *“Do you support the inclusion of a glossary of terms?”*. 100% agreement.

- **BC: Introduced discussion on number of items and whether general good practice items should be included**
- [note from the research team] General good practice items are recommendations applying to any types of study and not involving any AI specific components, like for example “state the study objectives”.
- **GC:** Shouldn’t aim for specific number, maybe room to merge some. Around 60 would probably be too much. Checklists are minimal reporting so some aspirational issues maybe better suited to E&E document.
- **BM:** Agrees probably no specific length. Wider point that word limits shouldn’t be a constraint. Should keep in generic good practice items.
- **PMC:** Danger that generic item inclusion drowns out importance to AI-specific items. Keep them in but make them clearly separate.
- **SS:** Remove generic items, aim for total count closer to 30 with only AI specific items. Smaller checklist more likely to be adopted. Potentially move generic items to different section.
- **BV:** Many generic items had very high rate of Delphi response suggesting inclusion. So perhaps question of where they go and how presented rather than whether or not to include them.
- **XL:** Asked whether scope pre-Delphi was to include generic items and whether some of the generic items have flowed through the process incidentally. **GC** agreed with this and sticking to AI specific items.
- **PW:** Maybe 3 categories: generic, AI specific, mixture.
- **BC:** Summarised that meeting consensus seems to be separation of AI specific items from generic items.
- **Vote held** -> *“In general, we should focus our list on AI specific items. Other (generic) items would be presented separately”*. 100% agreement.

- **BC introduced discussion for item 1a**
- **BC:** Not enthusiastic about using the word 'mention'. Change to 'specifying' or 'describing'.
- **SS:** Asked if we should decide on whether items are AI-specific (i.e. core) or generic (i.e. general good practice). **PMc** agreed with this.
- **XL:** Agrees with 'specifying'. Decision support system needs to be included in glossary and defined more clearly in E&E.
- **LM:** 'Early-stage' and 'formative' needs definition. Asked if these are the same. **BV** replied that aim was to not be too prescriptive hence both terms used. **PMc** suggested replace 'or' with '/' so the implication is that it isn't mutually exclusive.
- **GC:** Should be an AI specific item. It literally says AI. **XL, CR, DT, MM** agreed with this.
- **DT** also suggested more clarity between hierarchy of terms AI vs ML. And difficulty of definition of early-stage. **BC** mentioned that conversation could be addressed by the glossary.
- **LM:** The wording as it is doesn't really mandate mentioning AI, which is very important for indexing. A risk is that developer put in the title the name of their specific algorithm rather than mentioning it is AI/ML, which would not be helpful. The mention of AI/ML should be made clearer.
- **MM:** The specific way in which a decision support system is addressing a clinical problem is crucial to evaluation and need to be highlighted as well.
- **XL:** 3 things in this item. Asking about early stage, AI and clinical problem. This may affect adherence as people may only adhere to one aspect of an item to 'check' it off. Should be reworded to specify a + b + c.  
**DT** [via chat] agrees to disentangle it into bullet points  
**BC** suggested this could be separated by sub-items in the wording e.g. a, b, c.
- **DT:** "early-phase" and "formative" should be defined more clearly. **BC:** include these words in the glossary
- **WW:** Unclear on whether early-stage corresponds to e.g. phases in clinical trials.
- **PMc:** A compromise could be to agree on the general principle during the meeting and maybe the glossary work should be done offline.
- **PW:** These 3 are 'AND' statements rather than 'OR' so maybe clarify in that way.
- **SS:** Field is advancing too rapidly so better not to use words early-stage/formative at all. Use 'staging' so is more future proofed. **LM** agrees.
- **BV:** Wording has already gone through several rounds and in the interests of meeting logistics/time constraints better to avoid major changes unless critical at this point. The concept of "early-stage" is defining the scope of DECIDE-AI so should be kept in.
- **BC:** Summarised as 'early-stage' and "formative" issues will be left to glossary work by research team.

## Annex VI-2

- **Final wording:** “Identify the study as early-stage / formative clinical evaluation of an artificial intelligence or machine learning based decision support system, specifying the clinical problem addressed.”
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 93% for main list, 7% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 2**
- **BC:** Asked what people thought about the phrase 'state of the art'.
- **BG:** Define 'state of the art' in glossary. The intended definition for the term was probably more related to current existing practice. **BV** agreed with this re: current standard practice.
- **XL:** Asked what the purpose for this item is: get authors to compare to the most relevant comparator or to the best existing practice.
- **PMc:** Agrees with **BV**. Not about comparator given the early stage of this study.
- **PW:** Sounds like item is about not only standard practice but also usual care setting.
- **JO:** 'State of the art' has meaning in the regulatory world – e.g. manufactured according to standard best practice of the industry. If this isn't the intention then should be rephrased.
- **CR:** 'Standard of care' has specific legal meaning e.g. reasonableness standard. **MM** agreed and suggested standard practice would be better to avoid confusion.
- General discussion (**BC, PMc, BV**) on voting for items to be 3 way (include main list, include supplementary list, exclude completely). If >80% on any include option then gets included, majority between include options determines which list it goes into. However, after discussion consensus changed to 2 vote system. First vote on straight include/exclude (needs >= 80% of non absenting votes to be included). Next vote on which list -> simple majority determines which list.
- **Final wording:** "Describe the target clinical problem and medical condition, including the current standard practice, and the target patient population"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 71% for main list, 29% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 3**
- **XL:** Change wording to just potential impact.
- **PW:** Change wording to 'potential impact including patient outcomes'. Is "planned integration" referring to where or how?
- **PMc:** Agrees with **PW**. Many studies at this stage will be too small or underpowered to detect patient outcomes as opposed to process measures. 'Intended use' now has a formal definition under new EU rules – must have specific evidence for any claim.
- **JO:** MHRA says evidence should support intended use/purpose - has legal definition but seems reasonable to use in this context. "Intend use" and "intended purpose" are used interchangeably.
- **Final wording:** "Describe the intended use of the algorithm, its planned integration in the care pathway and the potential impact, including patient outcomes, it intends to achieve"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 86% for main list, 14% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 4**
- **JO:** There is international consensus around how software is described, although exact denomination differs for US vs. EU. Supports inclusion of the item. Regulatory aspect is helpful context for interpreting the scientific side of the algorithm. Could change the wording to “proposed device classification” or “pending regulatory approval”.
- **MM:** Important to state if regulatory approval has been received. Rest of the item is generic. Add “if applicable” to the part on regulatory approval.
- **CR:** No issue with scientific and regulatory both being together in the item. Many trials are made both for regulatory purposes and to demonstrate utility. Also feels item is generic.
- **PMc:** Dislikes mixing of regulatory and scientific issues. Hard to focus on both at the same time. Agrees item is generic.
- **BV:** Could remove the wording within brackets, potentially even exclude whole item. Refers to comments from Delphi participants about regulatory process evolving rapidly and that information about regulatory status might be outdated by the time of publication.
- **MM:** could be the right place to mention staging as previously suggested by **SS**.
- **PMc:** It sounds more like a risk classification. We don’t have a specific staging system for this field of research at the moment (although many similarities with IDEAL).
- **JO:** regulatory context is helpful information, especially if you are looking into using the algorithm in clinical context (can you legally use it or not?). Would favour splitting it in two items if included.
- **No change to wording:** “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”
- **Vote held** -> 60% for include, 40% for exclude.
- **Item EXCLUDED**
- **WW:** Asked if there is anywhere else algorithm will be described, as this is important to have in the publication. **BV** answered yes, in item 9.



- **BC introduced discussion for item 5**
- No comments from the group.
- **No change to wording:** “State the study objectives”
- **Vote held** -> 86% for include, 14% for exclude.
- **Vote held** -> 43% for main list, 57% for supplementary list.
- **Item INCLUDED in SUPPLEMENTARY LIST**

- **BC introduced discussion for item 6a**
- **MM:** General debate in the field about ethics provenance of ML work e.g. is it considered research or quality improvement. Favours item inclusion to make the process underlying the seeking of ethics approval more transparent.
- **SD:** Falls under standard practice. Feels that should be done in any case for journal to accept it. Should be included in the supplementary list.
- **WW:** Some early-stage studies might be preliminary and not registered in advance. Might not be able to make this mandatory. Most institutions would not let you do any research without ethics approval.
- **BC:** wording use “any”, so it is not mandatory but ask to reference it if there is one. **BV** confirms.
- **No change to wording:** “Provide a reference to any study protocol, study registration number and ethics approval”
- **Vote held** -> 86% for include, 14% for exclude.
- **Vote held** -> 20% for main list, 80% for supplementary list.
- **Item INCLUDED in SUPPLEMENTARY LIST**

## Annex VI-2

- **BC introduced discussion for item 6b**
- No comments from the group
- **No change to wording:** “State what measures were taken to protect patient privacy and data security”
- **Vote held** -> 7% for include, 93% for exclude.
- **Item EXCLUDED**

- **BC introduced discussion for item 7**
- **XL:** Item isn't vague but a bit generic and belongs in supplementary list.
- **BM:** Unless there is a list of study designs to choose from then it doesn't add much. Will be described later in the methods in any case and probably should be excluded.
- **GC:** Not very informative as it stands. If included then needs more detail.
- **BV:** Acknowledged thin line between reporting and methodology but asked if this could be an opportunity to help with the latter. The E&E section could give methodology pointers about the study design at this stage.
- **PMc:** No current defined stages so unclear what authors should describe for this item.
- **No change to wording:** "Describe the study design"
- **Vote held** -> 27% for include, 73% for exclude.
- **Item EXCLUDED**
- **GC and PMc** [via chat]: it would be useful to mention study design in the explanatory paper though. Seems odd not to have anything on it at all. All studies have a 'design'.

- **BC introduced discussion for items 8a and 8b**
- No objections to removing the word 'precisely'.
- **MM**: Feels recruitment is the right word, would not change to 'included to the study'. **GC** agrees.
- **XL**: Feels wording should expand to 'at both the patient and data level'. **BV** agrees.
- **BV**: 'Intended number of patients' is not meant to be a comment on sample size statistics. Likely to be underpowered anyway. Aim of the 'intended number' is for authors to justify why a specific number was chosen (i.e. was it just a convenience sample) rather than asking them to show detailed statistical calculations.
- **GC**: Asked if 'participants' should replace 'patients'.
- **BV** said reason for 'patients' was to highlight the distinction between patients and users of the decision support system.
- **PW** prefers 'participant'. **BC**: If patient representative prefer this word, this should be a strong argument in favour of using it.
- **XL**: this item is about the people whose data are used as input to the algorithm. There is another item about users coming afterwards.
- **MM** says there are 3 different populations: participants (those consented), patients and users (who may not be consented). Should focus on category for inclusion/exclusion criteria will have an impact on evaluation. **BC** says very important to add whatever terms are finally chosen into the glossary, though feels that patient is clearer.
- **PW**: There is a difference between patient whose data is used vs. patient involved in using the decision tool.
- **LM** says user vs. participant/patient is an important distinction.
- **MM**: Patients whose data feed into the algorithm might not always be (consented) participants (e.g. when routinely collected data are used). "Participant" connote a voluntary (consented) participation in the study. [via chat] users could also be consented to participate in the study.
- **BG** said a patient could also be a user if they are also involved with the decision support tool along with e.g. a clinician.
- **LM**: agrees with **BG**. The user part is very important, especially knowing if the users recruited during the evaluation represent the intended users (e.g. not only medical students).
- **PMc** suggests combining items 8a and 8b into a general recruitment item and then describe who it encompasses (patients, clinicians, users participants etc.) as further defined in the glossary.
- **LM**: Patients can be involved at 3 different levels (as part of the training set, in clinic when the decision tool is actually used to make a decision about them, as users of the algorithm output), glossary definition is important. Suggest a single, merged, item on participant recruitment and then explain the different categories in the E&E.

## Annex VI-2

- SD [via chat]: echoing SS comment earlier on, it is important to build a list that will still be relevant in N years.
- **SS** suggests delaying vote to third day.
- Discussion ended with **BV** offering to return to final day of consensus meeting with clarified wording for the group. **Vote postponed to another day** (see minutes for day 3 of consensus meeting).
- **BC introduced discussion for item 8c**
- **SD**: Should include the item because often failure of decision tools is partially a failure to engage clinicians and teach them about the tool.
- **JO**: Difference in training can lead to difference in decision tool performance. Similarly, level of experience i.e. seniority can affect success of a tool and is important to mention.
- **PW**: Unclear about the phrasing re: 'attempts to'. **BC** suggests 'steps taken to familiarise...'
- **MM**: Need to specify when the training is happening. Is it prior to study or during?
- **Final wording**: "Describe steps taken to familiarise the users with the algorithm, including any training received prior to the study"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 92% for main list, 8% for supplementary list
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 9**
- **GC:** Likes the item. Unclear on what 'expected performance in silico' means. **BV** says it is about the performance showcased during testing in a very controlled dataset and seeing how this compare to performance in actual clinical settings. More detail could go into E&E.
- **PW:** Asks if the reference to patient population is more about the data or the actual population. **BV** says it is in essence about ensuring that algorithm wasn't trained on a different population to who it will be used on.
- **MM:** A lot of this information would be in previous publications. The current DECIDE-AI study would be mainly aimed at a clinical audience so e.g. thresholds re: false positive or negatives are more important to state rather than esoteric technical details.
- **XL:** Everything currently in the list should stay. However, it might be impracticable to provide all this information without the reader going back to previous publications. Version number might be difficult to trace but as a general point it is important to be able to tease apart the version being evaluated and the one described in previous literature. Could instead refer to it, for example, as a unique device identifier rather than version number but the principle is to pin authors down and get them to commit to an identifier of some sort. Pre-regulatory is a more difficult sphere. **JO** [via chat] Given the stage of development there may not be a UDI yet.
- **JO:** If a UDI is available then state it. Else if version number is available then state that. Most current algorithms don't have peer-reviewed in silico trials/publications, so the algorithm description cannot rely solely on previous publications.
- **DT:** Agrees with XL that very few teams are using version numbers and reporting them in publications. Wonders if we may be being too aspirational in our requirements for authors. Staying practical is also important to increase uptake.
- **SS:** Need to know what algorithm is being evaluated. The guidelines could request a short description of the algorithm, possibly in the form of a separate 'product description document' that could be citable. The main study needs to be focused on evaluation itself and not on a comprehensive description of the underlying product. Giving a long list of algorithm characteristics to describe could distract from this main focus.
- **GC and JO** [via chat]: there is also this paper by Mark Sendak et al that presents a 'Model Facts' sheet that we can look at (<https://www.nature.com/articles/s41746-020-0253-3>). **MM** [via chat]: concern that some groups are generating these sheets based only on in silico validation only, where the 'risks' etc are speculative as it hasn't been tested in humans.
- **DT:** agrees with **SS**. Is past performance really so relevant to the current evaluation? At this stage the clinical performance is what really matters. We should focus on the intended scope of DECIDE-AI and leave more detailed description of algorithm to other reporting guidelines e.g. STARD-AI.
- **BV:** We need a description of some sort of the algorithm so readers know the context. Bigger question is what level of detail is desirable in the item itself and what could go to the E&E. Could work on rewording if needed.
- **BC** suggests voting on inclusion and then a later decision on level of granularity.

## Annex VI-2

- **No change to wording yet:** “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.”
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 86% for main list, 14% for supplementary list
- **Item INCLUDED in MAIN LIST**
- Final wording for the item and level of granularity still to be decided by consensus group on another day, time permitting.



## DAY 2

---

- **BC opened and introduced the session**
- **Present:**
  - **BC** - Bruce Campbell - (**Chair, non-voting**)
  - **GC** - Gary Collins
  - **SD** - Spiros Denaxas
  - **BG** - Bart Geerts
  - **XL** - Xiao Liu
  - **BM** - Bilal Mateen
  - **PM** - Piyush Mathur
  - **MM** - Melissa McCradden
  - **PMc** - Peter McCulloch
  - **LM** - Lauren Morgan
  - **JO** - Johan Ordish
  - **SS** - Suchi Saria
  - **DT** - Daniel Ting
  - **BV** - Baptiste Vasey
  - **WW** - Wim Weber
  - **PW** - Peter Wheatstone
- **Apologies:**
  - **CR** - Campbell Rogers
  - **SD** - Spiros Denaxas (from 14:00 to 15:00 BST)
- **BC:** Plan for the session is to go through as many items as possible. Ideally want to spend less time on green items today as (i) the green items had >80% consensus through both Delphi rounds, (ii) the green items did not receive objections from more than 2 of the 15 stakeholder groups, (iii) the research team has given the consensus group >2 weeks to offer comments prior to this meeting. No objections from the group to this approach proposed by **BC**.
- **BC:** Asked group if for high consensus items they want a vote at all about inclusion. **BV** mentioned that the steering group decided prior on the inclusion criteria and that an inclusion vote was necessary to include an item according to the procedural rules.

- **BC introduced discussion for item 10a**
- **BV:** A Delphi participant had suggested moving expected data input availability to item 10e. No objections from the group.
- **SS:** Suggested rewording to mention description of settings being in the current study. **PMc** disagreed as an author will be going through the checklist in order and so the context should be clear from the position of the item in the methods section.
- **BV:** Wording changed from 'was tested' to 'was evaluated' due to suggestion from **SS**.
- **Final wording:** "Describe the settings in which the algorithm was evaluated, including which additional clinical information (i.e. not provided by the algorithm) was accessible to the users to interpret or put into context the output of the algorithm"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 87% for main list, 13% for supplementary list
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for items 10b and 10c**
- **XL:** Feels implemented is more confusing than deployed. **MM** disagreed and felt deployment is not a term that clinicians would be keen on. **BC** suggested using the word 'evaluated' which also has the benefit of consistency with wording in item 10a.
- **BV:** Supportive of the Delphi participant's proposed change to "how the final clinical decision was made", as it includes shared decision making and conflict resolution. Proposed merging this item with 10c.
- **PM:** *Who* made the decision vs. *how* the decision is made is an important distinction. [via chat] Who, how and when are all important parts. **BC** suggests wording "how the final clinical decision is made and by whom".
- **LM** felt this might be too much for one item (in essence, who, how and when). **BV** replied that the who is dealt with by item 8b. This item is about *how* and *when*. **BC:** There is a difference between the use of the algorithm and the decision made.
- **BG:** Over time within the same hospital admission, different people may make different decisions at different time points as information changes. Makes the item complex. **MM** asked the group if these points in items 10b and 10c would fall into similar sentences when reported; if so, then might make sense to merge items together.
- **SD:** Very context dependant, the final decision might be presented to the clinician alone or by the clinician to the patient. In early evaluation, clinicians would most of the time make the final decision. **PW** said in this context we probably mean 'recommendation' rather than 'decision'. **BC** suggested that 'decision' adds nuance to separate this aspect from the algorithm recommendation.
- **BC** suggested wording as: "Describe the clinical workflow/pathway in which the algorithm was evaluated, the timing of its use, how the final decision was reached and by whom". **PM** agreed with this. The settings might also influence how the decision is made (e.g. smart watches). **LM** suggested removing the word 'clinical' from 'clinical decision', to encompass more type of AI applications. No other objections.
- **Final wording:** "Describe the clinical workflow/pathway in which the algorithm was evaluated, the timing of its use, and how the final decision was reached and by whom"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 100% for main list, 0% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 10d**
- **LM:** Unlikely (even concerning) to have IT integration prior to study if early-stage. Advocates item exclusion. **PW** agrees.
- **XL:** Two points. Is this item aimed at replicability and does the IT setup have implications on workflow and findings of the study. Also feels that algorithmic thresholds belong elsewhere.
- **BV:** Early-stage implementation may be local rather than integrated in any case. Algorithmic thresholds could be moved to the rest of the algorithm description in item 9.
- **BM:** Early-stage challenges of technical implementation might be useful for readers to know about. Suggests adding 'technical implementation' to newly merged item 10b/10c. **BC** and **PMc** disagreed as might make the super item too complex and that replicability does not mean ensuring this level of technical/granular detail. This is about scientific reporting, not audit. Moreover, if the developers intend to demonstrate some form of generalisability, the algorithm performance evaluation should not be dependent on local IT integration.
- **SS** was unsure how much detail is implied by this item.
- **No change to wording:** "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."
- **Vote held** -> 53% for include, 40% for exclude, 7% abstained.
- **Item EXCLUDED**

- **BC introduced discussion for item 10e**
- **BV:** Suggested wording change (expected availability of the data) from the offline discussion might be better placed in results rather than in this item.
- **MM:** Asked if the data in this item refers to user or participant. **BV** clarified that this relates to source data for the algorithm rather than users. Item 8a and 8b are about the inclusion/exclusion criteria, this item is about the type of data used and their handling. Will make it clear in the E&E section. **BC** suggested that clarification be included in the glossary of terms.
- **No change to wording:** “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”
- **Vote held** -> 93% for include, 7% for exclude.
- **Vote held** -> 87% for main list, 13% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 10f**
- **BV:** Mentioned that some Delphi participants suggested showing images of the display for this item. **LM** mentioned that in a recent systematic review of decision support systems the higher quality papers anecdotally did seem to include images of their display interface. Having an image is really useful to understand the display. **GC** had no intrinsic objections to this and suggested adding a mention to the end of the item e.g. “an image may be helpful” which is similar to the style used in other reporting guidelines e.g. PRISMA.
- **Final wording:** “Describe the algorithm outputs and how they were presented to the users (an image may be useful).”
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 100% for main list, 0% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 11a**
- **BC:** Suggests that adding 'if any' after secondary outcomes is probably unnecessary. No objection from the group.
- **No change to wording:** "Specify the primary and secondary outcomes measured"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 36% for main list, 64% for supplementary list.
- **Item INCLUDED in SUPPLEMENTARY LIST**

- **BC introduced discussion for item 11b**
- **PW:** Important to define what we mean by error.
- **DT:** Unclear what error metric we want reported: i.e. is it false positive/negatives or error rate.
- **XL:** Authors should have the flexibility to define the errors and which they decide to explore. **MM** agreed.
- **PMc:** Distinction between AI that is e.g. diagnostic and says “it’s this” vs decision support which says “do this”. The former has a potential gold standard whereas the latter does not (hence making it almost impossible to categorise things as errors with certainty). **BC** This difficulty is answered by the item. Felt this item is open enough to allow authors to address this issue.
- **XL:** also left open the definition of errors in her personal research work. Researchers can describe it as diagnostic errors, wrong decision making, or any instance where the recommendation does not bring patient benefits.
- **DT:** Is worried that leaving it to authors to define leaves it open to them artificially framing the study in a better light. **BC** mentioned that the aim in this work is reporting rather than methodology.
- **BV:** Raised point about **XL** suggestion for slight rephrase to ‘algorithm error’. No objections from the group.
- **Final wording:** “Describe how algorithm errors were defined and how they were identified”
- **Vote held** -> 93% for include, 7% for exclude.
- **Vote held** -> 93% for main list, 7% for supplementary list
- **Item INCLUDED in MAIN LIST**
- Technical issue with voting in that not all 15 showing. Test run of voting system showed no problems, unclear what occurred in the initial instance.



- **BC introduced discussion for item 12**
- **PMc:** Disagrees with removal of 'pre-specified' for the subgroup analyses. **GC** disagreed with this as pre-specified does not necessarily mean it was the right method and makes the item wording neater. Suggests 'describe the statistical methods by which the primary and secondary outcomes were analysed...'.  
**PM:** Suggested keeping the second 'pre-specified' in (the one preceding additional analyses). **PMc** and **GC** both agreed with this.
- **GC:** This would be a good place to add more detail in the E&E document.
- **Final wording:** "Describe the statistical methods by which the primary and secondary outcomes were analysed, as well as any pre-specified additional analyses, including subgroup analyses and their rationale."
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 64% for main list, 36% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for items 13a and 13b**
- **JO:** Would agree to merging of items 13a and 13b in principle.
- **GC:** Asked what 'preclinically' meant in this context, felt if from another study then should be in intro rather than methods. **BV** replied that it was based on AMLAS safety framework from University of York and **JO** agreed with this as someone who works on AMLAS with the team at York.
- **GC:** Asked if this should be moved to the algorithm item i.e. item 9. **XL** felt that best to leave this item flexible so that authors could define as they see fit. Authors doing the thinking is an important benefit already. Wondered if 'risk' was a better term than 'safety'. **JO** felt that the 13b part would not merge well with item 9.
- **PMc:** Safety depending on what happens with the output of the algorithm (regardless of whether it is necessarily accurate or not) and hence very context dependant. Will need a safety item (also very important from a regulatory perspective). There are methods to predict the level of risk of an application.
- **PM:** Suggested even broader wording along the lines of "define the safety requirements and how these were evaluated".
- **Final wording** to be re-presented on day 3 of the consensus meeting as a merged item including 13a and 13b components.
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 92% for main list, 8% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 14**
- **LM:** Human factors (HF) work is important and central to scope of DECIDE-AI studies. Excluding this item in the methods section would have knock on effect on HF items in results section (though there may be room to merge some of those).
- **MM:** Asked whether 'if applicable' should be added, in case some research group don't have access to HF expertise. **LM** felt that this is a core component of new technology evaluation and considering this, adding "if applicable" to the HF item would be like adding it to the item on statistical analysis or outcomes. **PM** felt strongly that HF is a very important aspect of these studies, too often ignored.
- A partial vote occurred (as discussion continued during the voting process) with results showing 80% include, 15% exclude and one abstention i.e. item very on the margin for whether it is included or not.
- **SS:** Every tool should have HF work but it may not be present as too much to include in a single manuscript. This would be a separate study and we could do a whole checklist on just HF aspects. **LM** strongly disagreed and stated HF work probably more important than e.g. clinical outcomes in early stage studies.
- **SS:** HF should be advocated more, but it is not practicable to mandate a full description of HF methodology and results in the same studies describing early-stage clinical utility. **GC** and **PMc** also strongly disagreed and support inclusion of HF. [GC posted the link to the original DECIDE-AI correspondence in the chat.] The whole online premise of DECIDE-AI was about HF, this cannot be ignored at this stage or would defeat the purpose of the guidelines. While granularity of detail required is debatable, both felt that presence should be mandatory. HF and statistical elements in the same study perfectly compatible.
- **LM:** Could remove 'human factors' at the end of the sentence. No objections from the group.
- **Final wording:** "Describe the human factors tools, methods or frameworks used, the use cases considered and the users involved"
- **Vote held** -> 79% for include, 14% for exclude, 7% abstention -> 80% threshold is reached as this threshold only apply to non-blank votes, so 84% voted in effect to include consistent with the agreed voting practice. Item is therefore included. Recognising the closeness of the vote, **BV** happy to discuss offline how the item is elaborated on in the E&E document for anyone with strong feelings and to reflect the minority's point of view.
- **Vote held** -> 86% for main list, 14% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 15**
- **MM:** Inclined not to be too prescriptive even though it is important. Suggested possible merger with item 16.
- **BV:** Both items 15 and 16 received fairly low consensus among Delphi participants. Could merge them and talk more broadly about stakeholder involvement.
- **BC:** Emphasised that UK government position is for patient involvement in all aspects of healthcare and asked **JO** for his views as a member of the MHRA. **JO** agreed though conceded it may not necessarily have to be at this stage. **BC:** FDA is also increasing the focus on patient involvement.
- **PW:** Would be open to merging items 15 and 16. However, mention of 'any stakeholder' waters down the patient and ethics focused element of this item.
- **GC:** Thinks both items will be on supplementary list if brought through so might be better not to merge. Separate reporting guidance exists for PPI work (GRIPP2).
- **MM:** suggested (acknowledging contradiction with her previous statement) that merger might confuse two different issues (i.e. item 16 is more about algorithmic bias/fairness and therefore more of a specific methods issue compared to broader topic of patient engagement).
- Technical issue with internet connection to **BC** so discussion on this item paused and moved onto other items with **GC** as temporary chair. Group returned to this discussion once **BC** reconnected to the meeting.
- **PW:** It is tokenistic as currently worded unless it asks authors to mention 'how' they were involved. No objections to this change from the group.
- **Final wording:** "State how patients were involved in any aspect of the study design, conduct or in the development of the research question or outcome measures"
- **Vote held** -> 86% for include, 7% for exclude, 7% abstention.
- **Vote held** -> 7% for main list, 93% for supplementary list.
- **Item INCLUDED in SUPPLEMENTARY LIST**

- **BC introduced discussion for item 16**
- **MM:** Main focus of the item should be algorithmic fairness and steps taken to ensure it. For example, if bias against certain patient/user groups were observed in the development phase, researchers might apply a correction factor to ensure equitable allocation to services, which would impact on the performance [possibly negatively but increasing fairness] If such correction is not possible, then it is important to know how this was communicated to clinicians. Knowing about these factors is important for reproducibility as they are likely to impact performance. It would be good to be a bit more specific about what is expected (e.g. specifying algorithmic fairness). Could add more guidance in the E&E document.
- **BC:** Suggested adding wording about algorithmic fairness to address this. No objections raised by the group.
- **XL:** Destination between main and supplementary list depends on how big a focus 'algorithmic fairness' forms for the study in question. If a big part, then more likely to be important for the AI specific list. **MM** replied that currently algorithmic fairness is not well reported.
- **GC:** considerations about specific vulnerable subgroups could be address in item 22 (subgroup analysis). **XL:** Should not reduce this to a simple sub-group analysis in the results. Subgroup analysis is one way of evaluating the ethical aspects of an algorithm but not sufficient. **MM** agreed with **XL**.
- **SS:** Too much to expect full evaluation of e.g. algorithmic fairness, human factors, clinical outcomes all in the same manuscript. A full evaluation of a tool should be a collection of paper. **BC** felt that shouldn't necessarily detract from the purpose of this guideline. **SS:** Our list works well if listing all elements that need to be reported for the overall evaluation of a tool, but would be too much for a checklist that all papers need to address. **MM:** agreed that this is not about doing a full fairness evaluation, simply stating what researchers did in term of correction for algorithmic fairness (can be done at different stages of the data processing).
- **LM:** Shouldn't need to narrow down to allow everything to fit into 1 manuscript, DECIDE-AI should encourage referencing sideways. **PMc** agreed with **LM** and mentioned in his role as an editor for some journals that there is discretion on e.g. word counts and level of adherence to guidelines. **GC** agreed.
- **GC:** Given the topical nature of algorithmic fairness in AI, it should be added to the item.
- General group consensus to add wording on algorithmic fairness.
- **SS:** Are we advertising this list as a general set of guidelines for the evaluation of AI tools or a checklist were every single item need to be checked? Important to inform her vote as, in the latter case, there is a need to prioritise. **PMc:** as previously stated guidelines are guidance not laws. Even if checklist, nobody will be bound by them. **GC** agreed and stated that even the most famous reporting guidelines (e.g. original CONSORT) don't garner full adherence. **SS** disagreed.
- **BC:** the way the guidelines will be presented is a matter of discussion for a later stage and will be discussed when drafting the manuscript. Proposed moving to a vote.
- **JO:** current wording say "describe any" so is not mandating such description anyway. **GC** agreed.
- **Final wording:** "Describe any ethics methodology, consultation or involvement during the design or implementation of the study, such as algorithmic fairness".

## Annex VI-2

- **Vote held** -> 86% for include, 7% for exclude, 7% abstention.
- **Vote held** -> 71% for main list, 29% for supplementary list.
- **Item INCLUDED in MAIN LIST**
- **GC [Chair ad interim] introduced discussion for item 17a and 17b**
- **LM:** should be discussed once the agreement has been reached about 8a and 8b. **XL:** these are slightly different topic as some of the data mentioned here might come from group who didn't necessarily chose to participate.
- **BV:** Asked group whether we should keep the brackets and their content or move it to the E&E document. **XL** felt better to leave it flexible/open (moving the bracket's content to the E&E).
- Decision to come back to this item along the lines of item 8a and 8b when wording clarified, on day 3 of the consensus meeting.

- **GC [Chair ad interim] introduced discussion for item 18a**
- **GC:** Some of the terms probably need more clarification and the wording is a bit clunky. **BV** agreed that this could be covered in the glossary / E&E. Proposed to remove the brackets. No objections from the group to removing the words in brackets.
- **Final wording:** “Report on the user exposure to the algorithm, on the number of instances the algorithm was used and on the users’ adherence to the intended implementation”
- **Vote held** -> 87% for include, 7% for exclude, 7% abstention.
- **Vote held** -> 80% for main list, 20% for supplementary list.
- **Item INCLUDED in MAIN LIST**
- **Technical issue for BC re: internet connection resolved and BC took over again as chair from this point going back to items 15 and 16 as well.**

- **BC introduced discussion for item 18b**
- No comments from the group.
- **No change to wording:** “Report changes caused by the algorithm to the clinical workflow, if any”
- **Vote held** -> 71% for include, 29% for exclude.
- **Item EXCLUDED** -> though later re-vote and ultimately included during day 3 of consensus meeting. See explanations and justification in meeting minutes for day 3.



- **BC introduced discussion for item 19**
- **BC:** Noted a suggestion from offline exchanges to add 'and the reasons for' after modifications
- **LM:** Wording could be condensed to 'during the study'
- **XL:** Asked if the presence of this item inherently allows the scope of DECIDE-AI to include studies where the algorithm is permitted to change during the course of the study. **LM** gave example of study in which version changes were reported and described. **PMc** replied that this was based on IDEAL stage 2a and that changes may occur anyway during early-stage studies (as frequently occurs in non-AI surgery studies, despite unfortunately being very rarely reported, hence concealing important information and increasing the risk of people repeating mistakes). This item ensures that at least such changes are transparently reported. **BV** added that iteration of the algorithm in the early stage may be useful, even desirable, and could occur frequently. The difference with later, summative, evaluation like RCTs (in which a change in the intervention is indeed problematic) must be clearly stated in the E&E. Reporting on iterative improvements in early-stage is also one of the key aspects of the DECIDE-AI, but might take time to be accepted.
- **Final wording:** "Report any changes made to the algorithm or its hardware platform during the study. Report the timing of these modifications, the rationale for them, and the change in outcomes observed after each of them".
- **Vote held** -> 93% for include, 0% for exclude, 7% abstention.
- **Vote held** -> 93% for main list, 7% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 20a**
- **BV:** Asked if variation over time, whilst important, may be better placed in a more consistent item. At the moment, this is mentioned in several different places (19, 20a, 23b, 23d).
- **MM:** is variation over time part of a planned analysis?
- **LM** unclear on the wording of the question (“algorithm-assisted users”). **BV:** this is a legacy of former iteration of the list in which different types of outcomes were considered.
- **LM:** part about variation in time is duplicative of e.g. item 23d.
- **GC:** time was not mentioned in the methods section. Suggested rewording to “what are the results of the pre-specified primary and secondary outcomes”
- **BC:** Suggested rewording of the item and then returning to it on day 3 of consensus meeting.
- **BV** will return with new wording for tomorrow.

## DAY 3

---

- **BC opened and introduced the session**
- **Present:**
  - **BC** - Bruce Campbell - (**Chair, non-voting**)
  - **GC** - Gary Collins
  - **SD** - Spiros Denaxas
  - **BG** - Bart Geerts
  - **XL** - Xiao Liu
  - **BM** - Bilal Mateen
  - **PM** - Piyush Mathur
  - **MM** - Melissa McCradden
  - **PMc** - Peter McCulloch
  - **LM** - Lauren Morgan
  - **JO** - Johan Ordish
  - **SS** - Suchi Saria
  - **DT** - Daniel Ting
  - **BV** - Baptiste Vasey
  - **WW** - Wim Weber
  - **PW** - Peter Wheatstone
- **Apologies:**
  - **CR** - Campbell Rogers
  - **SD** - Spiros Denaxas (from 14:00 to 15:00 BST)
- **BC:** Plan for the session is to go through as many items as possible. Ideally want to spend less time on green items today as (i) the green items had >80% consensus through both Delphi rounds, (ii) the green items did not receive objections from more than 2 of the 15 stakeholder groups, (iii) the research team has given the consensus group >2 weeks to offer comments prior to this meeting. No objections from the group to this approach proposed by **BC**.
- **BC:** Asked group if for high consensus items they want a vote at all about inclusion. **BV** mentioned that the steering group decided prior on the inclusion criteria and that an inclusion vote was necessary to include an item according to the procedural rules.

- **BC introduced discussion for item 10a**
- **BV:** A Delphi participant had suggested moving expected data input availability to item 10e. No objections from the group.
- **SS:** Suggested rewording to mention description of settings being in the current study. **PMc** disagreed as an author will be going through the checklist in order and so the context should be clear from the position of the item in the methods section.
- **BV:** Wording changed from 'was tested' to 'was evaluated' due to suggestion from **SS**.
- **Final wording:** "Describe the settings in which the algorithm was evaluated, including which additional clinical information (i.e. not provided by the algorithm) was accessible to the users to interpret or put into context the output of the algorithm"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 87% for main list, 13% for supplementary list
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for items 10b and 10c**
- **XL:** Feels implemented is more confusing than deployed. **MM** disagreed and felt deployment is not a term that clinicians would be keen on. **BC** suggested using the word 'evaluated' which also has the benefit of consistency with wording in item 10a.
- **BV:** Supportive of the Delphi participant's proposed change to "how the final clinical decision was made", as it includes shared decision making and conflict resolution. Proposed merging this item with 10c.
- **PM:** *Who* made the decision vs. *how* the decision is made is an important distinction. [via chat] Who, how and when are all important parts. **BC** suggests wording "how the final clinical decision is made and by whom".
- **LM** felt this might be too much for one item (in essence, who, how and when). **BV** replied that the who is dealt with by item 8b. This item is about *how* and *when*. **BC:** There is a difference between the use of the algorithm and the decision made.
- **BG:** Over time within the same hospital admission, different people may make different decisions at different time points as information changes. Makes the item complex. **MM** asked the group if these points in items 10b and 10c would fall into similar sentences when reported; if so, then might make sense to merge items together.
- **SD:** Very context dependant, the final decision might be presented to the clinician alone or by the clinician to the patient. In early evaluation, clinicians would most of the time make the final decision. **PW** said in this context we probably mean 'recommendation' rather than 'decision'. **BC** suggested that 'decision' adds nuance to separate this aspect from the algorithm recommendation.
- **BC** suggested wording as: "Describe the clinical workflow/pathway in which the algorithm was evaluated, the timing of its use, how the final decision was reached and by whom". **PM** agreed with this. The settings might also influence how the decision is made (e.g. smart watches). **LM** suggested removing the word 'clinical' from 'clinical decision', to encompass more type of AI applications. No other objections.
- **Final wording:** "Describe the clinical workflow/pathway in which the algorithm was evaluated, the timing of its use, and how the final decision was reached and by whom"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 100% for main list, 0% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 10d**
- **LM:** Unlikely (even concerning) to have IT integration prior to study if early-stage. Advocates item exclusion. **PW** agrees.
- **XL:** Two points. Is this item aimed at replicability and does the IT setup have implications on workflow and findings of the study. Also feels that algorithmic thresholds belong elsewhere.
- **BV:** Early-stage implementation may be local rather than integrated in any case. Algorithmic thresholds could be moved to the rest of the algorithm description in item 9.
- **BM:** Early-stage challenges of technical implementation might be useful for readers to know about. Suggests adding 'technical implementation' to newly merged item 10b/10c. **BC** and **PMc** disagreed as might make the super item too complex and that replicability does not mean ensuring this level of technical/granular detail. This is about scientific reporting, not audit. Moreover, if the developers intend to demonstrate some form of generalisability, the algorithm performance evaluation should not be dependent on local IT integration.
- **SS** was unsure how much detail is implied by this item.
- **No change to wording:** "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."
- **Vote held** -> 53% for include, 40% for exclude, 7% abstained.
- **Item EXCLUDED**

- **BC introduced discussion for item 10e**
- **BV:** Suggested wording change (expected availability of the data) from the offline discussion might be better placed in results rather than in this item.
- **MM:** Asked if the data in this item refers to user or participant. **BV** clarified that this relates to source data for the algorithm rather than users. Item 8a and 8b are about the inclusion/exclusion criteria, this item is about the type of data used and their handling. Will make it clear in the E&E section. **BC** suggested that clarification be included in the glossary of terms.
- **No change to wording:** “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”
- **Vote held** -> 93% for include, 7% for exclude.
- **Vote held** -> 87% for main list, 13% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 10f**
- **BV:** Mentioned that some Delphi participants suggested showing images of the display for this item. **LM** mentioned that in a recent systematic review of decision support systems the higher quality papers anecdotally did seem to include images of their display interface. Having an image is really useful to understand the display. **GC** had no intrinsic objections to this and suggested adding a mention to the end of the item e.g. “an image may be helpful” which is similar to the style used in other reporting guidelines e.g. PRISMA.
- **Final wording:** “Describe the algorithm outputs and how they were presented to the users (an image may be useful).”
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 100% for main list, 0% for supplementary list.
- **Item INCLUDED in MAIN LIST**



- **BC introduced discussion for item 11a**
- **BC:** Suggests that adding 'if any' after secondary outcomes is probably unnecessary. No objection from the group.
- **No change to wording:** "Specify the primary and secondary outcomes measured"
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 36% for main list, 64% for supplementary list.
- **Item INCLUDED in SUPPLEMENTARY LIST**

- **BC introduced discussion for item 11b**
- **PW:** Important to define what we mean by error.
- **DT:** Unclear what error metric we want reported: i.e. is it false positive/negatives or error rate.
- **XL:** Authors should have the flexibility to define the errors and which they decide to explore. **MM** agreed.
- **PMc:** Distinction between AI that is e.g. diagnostic and says “it’s this” vs decision support which says “do this”. The former has a potential gold standard whereas the latter does not (hence making it almost impossible to categorise things as errors with certainty). **BC** This difficulty is answered by the item. Felt this item is open enough to allow authors to address this issue.
- **XL:** also left open the definition of errors in her personal research work. Researchers can describe it as diagnostic errors, wrong decision making, or any instance where the recommendation does not bring patient benefits.
- **DT:** Is worried that leaving it to authors to define leaves it open to them artificially framing the study in a better light. **BC** mentioned that the aim in this work is reporting rather than methodology.
- **BV:** Raised point about **XL** suggestion for slight rephrase to ‘algorithm error’. No objections from the group.
- **Final wording:** “Describe how algorithm errors were defined and how they were identified”
- **Vote held** -> 93% for include, 7% for exclude.
- **Vote held** -> 93% for main list, 7% for supplementary list
- **Item INCLUDED in MAIN LIST**
- Technical issue with voting in that not all 15 showing. Test run of voting system showed no problems, unclear what occurred in the initial instance.

- **BC introduced discussion for item 12**
- **PMc:** Disagrees with removal of 'pre-specified' for the subgroup analyses. **GC** disagreed with this as pre-specified does not necessarily mean it was the right method and makes the item wording neater. Suggests 'describe the statistical methods by which the primary and secondary outcomes were analysed...'.  
**PM:** Suggested keeping the second 'pre-specified' in (the one preceding additional analyses). **PMc** and **GC** both agreed with this.
- **GC:** This would be a good place to add more detail in the E&E document.
- **Final wording:** "Describe the statistical methods by which the primary and secondary outcomes were analysed, as well as any pre-specified additional analyses, including subgroup analyses and their rationale."
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 64% for main list, 36% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for items 13a and 13b**
- **JO:** Would agree to merging of items 13a and 13b in principle.
- **GC:** Asked what 'preclinically' meant in this context, felt if from another study then should be in intro rather than methods. **BV** replied that it was based on AMLAS safety framework from University of York and **JO** agreed with this as someone who works on AMLAS with the team at York.
- **GC:** Asked if this should be moved to the algorithm item i.e. item 9. **XL** felt that best to leave this item flexible so that authors could define as they see fit. Authors doing the thinking is an important benefit already. Wondered if 'risk' was a better term than 'safety'. **JO** felt that the 13b part would not merge well with item 9.
- **PMc:** Safety depending on what happens with the output of the algorithm (regardless of whether it is necessarily accurate or not) and hence very context dependant. Will need a safety item (also very important from a regulatory perspective). There are methods to predict the level of risk of an application.
- **PM:** Suggested even broader wording along the lines of "define the safety requirements and how these were evaluated".
- **Final wording** to be re-presented on day 3 of the consensus meeting as a merged item including 13a and 13b components.
- **Vote held** -> 100% for include, 0% for exclude.
- **Vote held** -> 92% for main list, 8% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 14**
- **LM:** Human factors (HF) work is important and central to scope of DECIDE-AI studies. Excluding this item in the methods section would have knock on effect on HF items in results section (though there may be room to merge some of those).
- **MM:** Asked whether 'if applicable' should be added, in case some research group don't have access to HF expertise. **LM** felt that this is a core component of new technology evaluation and considering this, adding "if applicable" to the HF item would be like adding it to the item on statistical analysis or outcomes. **PM** felt strongly that HF is a very important aspect of these studies, too often ignored.
- A partial vote occurred (as discussion continued during the voting process) with results showing 80% include, 15% exclude and one abstention i.e. item very on the margin for whether it is included or not.
- **SS:** Every tool should have HF work but it may not be present as too much to include in a single manuscript. This would be a separate study and we could do a whole checklist on just HF aspects. **LM** strongly disagreed and stated HF work probably more important than e.g. clinical outcomes in early stage studies.
- **SS:** HF should be advocated more, but it is not practicable to mandate a full description of HF methodology and results in the same studies describing early-stage clinical utility. **GC** and **PMc** also strongly disagreed and support inclusion of HF. [GC posted the link to the original DECIDE-AI correspondence in the chat.] The whole online premise of DECIDE-AI was about HF, this cannot be ignored at this stage or would defeat the purpose of the guidelines. While granularity of detail required is debatable, both felt that presence should be mandatory. HF and statistical elements in the same study perfectly compatible.
- **LM:** Could remove 'human factors' at the end of the sentence. No objections from the group.
- **Final wording:** "Describe the human factors tools, methods or frameworks used, the use cases considered and the users involved"
- **Vote held** -> 79% for include, 14% for exclude, 7% abstention -> 80% threshold is reached as this threshold only apply to non-blank votes, so 84% voted in effect to include consistent with the agreed voting practice. Item is therefore included. Recognising the closeness of the vote, **BV** happy to discuss offline how the item is elaborated on in the E&E document for anyone with strong feelings and to reflect the minority's point of view.
- **Vote held** -> 86% for main list, 14% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 15**
- **MM:** Inclined not to be too prescriptive even though it is important. Suggested possible merger with item 16.
- **BV:** Both items 15 and 16 received fairly low consensus among Delphi participants. Could merge them and talk more broadly about stakeholder involvement.
- **BC:** Emphasised that UK government position is for patient involvement in all aspects of healthcare and asked **JO** for his views as a member of the MHRA. **JO** agreed though conceded it may not necessarily have to be at this stage. **BC:** FDA is also increasing the focus on patient involvement.
- **PW:** Would be open to merging items 15 and 16. However, mention of 'any stakeholder' waters down the patient and ethics focused element of this item.
- **GC:** Thinks both items will be on supplementary list if brought through so might be better not to merge. Separate reporting guidance exists for PPI work (GRIPP2).
- **MM:** suggested (acknowledging contradiction with her previous statement) that merger might confuse two different issues (i.e. item 16 is more about algorithmic bias/fairness and therefore more of a specific methods issue compared to broader topic of patient engagement).
- Technical issue with internet connection to **BC** so discussion on this item paused and moved onto other items with **GC** as temporary chair. Group returned to this discussion once **BC** reconnected to the meeting.
- **PW:** It is tokenistic as currently worded unless it asks authors to mention 'how' they were involved. No objections to this change from the group.
- **Final wording:** "State how patients were involved in any aspect of the study design, conduct or in the development of the research question or outcome measures"
- **Vote held** -> 86% for include, 7% for exclude, 7% abstention.
- **Vote held** -> 7% for main list, 93% for supplementary list.
- **Item INCLUDED in SUPPLEMENTARY LIST**

- **BC introduced discussion for item 16**
- **MM:** Main focus of the item should be algorithmic fairness and steps taken to ensure it. For example, if bias against certain patient/user groups were observed in the development phase, researchers might apply a correction factor to ensure equitable allocation to services, which would impact on the performance [possibly negatively but increasing fairness] If such correction is not possible, then it is important to know how this was communicated to clinicians. Knowing about these factors is important for reproducibility as they are likely to impact performance. It would be good to be a bit more specific about what is expected (e.g. specifying algorithmic fairness). Could add more guidance in the E&E document.
- **BC:** Suggested adding wording about algorithmic fairness to address this. No objections raised by the group.
- **XL:** Destination between main and supplementary list depends on how big a focus 'algorithmic fairness' forms for the study in question. If a big part, then more likely to be important for the AI specific list. **MM** replied that currently algorithmic fairness is not well reported.
- **GC:** considerations about specific vulnerable subgroups could be address in item 22 (subgroup analysis). **XL:** Should not reduce this to a simple sub-group analysis in the results. Subgroup analysis is one way of evaluating the ethical aspects of an algorithm but not sufficient. **MM** agreed with **XL**.
- **SS:** Too much to expect full evaluation of e.g. algorithmic fairness, human factors, clinical outcomes all in the same manuscript. A full evaluation of a tool should be a collection of paper. **BC** felt that shouldn't necessarily detract from the purpose of this guideline. **SS:** Our list works well if listing all elements that need to be reported for the overall evaluation of a tool, but would be too much for a checklist that all papers need to address. **MM:** agreed that this is not about doing a full fairness evaluation, simply stating what researchers did in term of correction for algorithmic fairness (can be done at different stages of the data processing).
- **LM:** Shouldn't need to narrow down to allow everything to fit into 1 manuscript, DECIDE-AI should encourage referencing sideways. **PMc** agreed with **LM** and mentioned in his role as an editor for some journals that there is discretion on e.g. word counts and level of adherence to guidelines. **GC** agreed.
- **GC:** Given the topical nature of algorithmic fairness in AI, it should be added to the item.
- General group consensus to add wording on algorithmic fairness.
- **SS:** Are we advertising this list as a general set of guidelines for the evaluation of AI tools or a checklist were every single item need to be checked? Important to inform her vote as, in the latter case, there is a need to prioritise. **PMc:** as previously stated guidelines are guidance not laws. Even if checklist, nobody will be bound by them. **GC** agreed and stated that even the most famous reporting guidelines (e.g. original CONSORT) don't garner full adherence. **SS** disagreed.
- **BC:** the way the guidelines will be presented is a matter of discussion for a later stage and will be discussed when drafting the manuscript. Proposed moving to a vote.
- **JO:** current wording say "describe any" so is not mandating such description anyway. **GC** agreed.
- **Final wording:** "Describe any ethics methodology, consultation or involvement during the design or implementation of the study, such as algorithmic fairness".

## Annex VI-2

- **Vote held** -> 86% for include, 7% for exclude, 7% abstention.
- **Vote held** -> 71% for main list, 29% for supplementary list.
- **Item INCLUDED in MAIN LIST**
- **GC [Chair ad interim] introduced discussion for item 17a and 17b**
- **LM:** should be discussed once the agreement has been reached about 8a and 8b. **XL:** these are slightly different topic as some of the data mentioned here might come from group who didn't necessarily chose to participate.
- **BV:** Asked group whether we should keep the brackets and their content or move it to the E&E document. **XL** felt better to leave it flexible/open (moving the bracket's content to the E&E).
- Decision to come back to this item along the lines of item 8a and 8b when wording clarified, on day 3 of the consensus meeting.



- **GC [Chair ad interim] introduced discussion for item 18a**
- **GC:** Some of the terms probably need more clarification and the wording is a bit clunky. **BV** agreed that this could be covered in the glossary / E&E. Proposed to remove the brackets. No objections from the group to removing the words in brackets.
- **Final wording:** “Report on the user exposure to the algorithm, on the number of instances the algorithm was used and on the users’ adherence to the intended implementation”
- **Vote held** -> 87% for include, 7% for exclude, 7% abstention.
- **Vote held** -> 80% for main list, 20% for supplementary list.
- **Item INCLUDED in MAIN LIST**
- **Technical issue for BC re: internet connection resolved and BC took over again as chair from this point going back to items 15 and 16 as well.**

- **BC introduced discussion for item 18b**
- No comments from the group.
- **No change to wording:** “Report changes caused by the algorithm to the clinical workflow, if any”
- **Vote held** -> 71% for include, 29% for exclude.
- **Item EXCLUDED** -> though later re-vote and ultimately included during day 3 of consensus meeting. See explanations and justification in meeting minutes for day 3.

- **BC introduced discussion for item 19**
- **BC:** Noted a suggestion from offline exchanges to add 'and the reasons for' after modifications
- **LM:** Wording could be condensed to 'during the study'
- **XL:** Asked if the presence of this item inherently allows the scope of DECIDE-AI to include studies where the algorithm is permitted to change during the course of the study. **LM** gave example of study in which version changes were reported and described. **PMc** replied that this was based on IDEAL stage 2a and that changes may occur anyway during early-stage studies (as frequently occurs in non-AI surgery studies, despite unfortunately being very rarely reported, hence concealing important information and increasing the risk of people repeating mistakes). This item ensures that at least such changes are transparently reported. **BV** added that iteration of the algorithm in the early stage may be useful, even desirable, and could occur frequently. The difference with later, summative, evaluation like RCTs (in which a change in the intervention is indeed problematic) must be clearly stated in the E&E. Reporting on iterative improvements in early-stage is also one of the key aspects of the DECIDE-AI, but might take time to be accepted.
- **Final wording:** "Report any changes made to the algorithm or its hardware platform during the study. Report the timing of these modifications, the rationale for them, and the change in outcomes observed after each of them".
- **Vote held** -> 93% for include, 0% for exclude, 7% abstention.
- **Vote held** -> 93% for main list, 7% for supplementary list.
- **Item INCLUDED in MAIN LIST**

- **BC introduced discussion for item 20a**
- **BV:** Asked if variation over time, whilst important, may be better placed in a more consistent item. At the moment, this is mentioned in several different places (19, 20a, 23b, 23d).
- **MM:** is variation over time part of a planned analysis?
- **LM** unclear on the wording of the question (“algorithm-assisted users”). **BV:** this is a legacy of former iteration of the list in which different types of outcomes were considered.
- **LM:** part about variation in time is duplicative of e.g. item 23d.
- **GC:** time was not mentioned in the methods section. Suggested rewording to “what are the results of the pre-specified primary and secondary outcomes”
- **BC:** Suggested rewording of the item and then returning to it on day 3 of consensus meeting.
- **BV** will return with new wording for tomorrow.