

Supplementary Table 1. Characteristics of the Delphi study and consensus meeting participants.

Participants	Delphi survey n=103 (%)	Consensus meeting n=31 (%)
Area of Expertise		
Healthcare professional	25(24)	5(16)
Methodologist/Statistician	20(19)	5(16)
Computer scientists	15(14)	3(9)
Industry representatives	11(10)	3(9)
Journal editors	10(9)	6(19)
Policy-makers	6(5)	1(3)
Informatics and healthcare delivery	5(4)	0(-)
Regulators	5(4)	2(6)
Patient advocates	5(4)	3(9)
Funders	4(3)	2(6)
Law and ethics	3(2)	1(3)
Other	14(13)	0(-)
Experience with clinical trials		
Trial design	49(47)	11(35)
Trial analysis	57(55)	11(35)
Trial reporting	52(50)	14(45)
Reviewing trials funding	42(40)	10(32)
Research ethics for trials	41(39)	11(35)
Advisory role for policy-makers or commissioning groups for clinical trials	26(25)	5(16)
Some theoretical knowledge but not direct experience	40(38)	7(22)
Additional experience in clinical trials	2(1)	1(3)
Experience with AI/ML		
Designing studies to validate AI/ML models	46(44)	11(35)
Developing AI/ML models	47(45)	9(29)
Reviewing AI/ML funding applications	44(42)	9(29)
Implementation of AI/ML in a clinical context	47(45)	7(22)
Some theoretical knowledge on AI/ML but not direct experience	43(41)	12(38)
Advising on transparency and reproducibility of AI/ML models	42(40)	12(38)
Advising on the ethical implications of AI/ML models	31(30)	6(19)
Additional experience in AI	10(9)	3(9)

AI (Artificial intelligence), ML (machine learning)

Participants could select multiple areas of expertise and multiple areas of experience with clinical trials and AI/ML.

Number of participants with expertise in clinical trials and AI/ML: healthcare professionals (n=21); methodologist/statistician (n=18); computer science (n=14); industry representatives (n=5); journal editors (n=9); policy-makers (n=5); informatics and healthcare delivery (n=5); regulators (n=2); patient advocate (n=1); funders (n=1); and law and ethics (n=1).

Supplementary Table 2. Consensus meeting notes and decisions for SPIRIT-AI and CONSORT-AI

	SPIRIT-AI	CONSORT-AI	SPIRIT-AI	CONSORT-AI	SPIRIT-AI	CONSORT-AI	Reasons for exclusion	Consensus meeting discussion notes	Final SPIRIT-AI item	Final CONSORT-AI item
	Delphi median score (IQR)	Delphi median score (IQR)	(%) INCLUDE	(%) EXCLUDE	(%) INCLUDE	(%) EXCLUDE				
Candidate items arising from Delphi Surveys										
Identify the intervention as an AI/machine learning intervention and specify the type of machine learning	8.0 (7.0-9.0)	8.0 (7.0-9.0)	94	6	94	6		AI may reach broader audience and it might be considered as a more sensitive term. Title should not be too lengthy. AI as opposed to ML may be easier and more accessible for clinicians and systematic reviewers; specification in the abstract. Umbrella term is useful in a situation of evolving terminology. Artificial intelligence and machine learning are useful but the architecture/model is not (consider different training datasets). Regulatory term "medical device". General terms are more useful from a long-term perspective.	Item 1(i) Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model	Item 1a,b (i) Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model
Specify the purpose of the AI intervention	8.0 (7.0-9.0)	8.0 (7.0-9.0)	90	10	87	13		Description should be harmonised with regulatory guidance. The specific use should be specified early on, but the intended use can evolve as the technology develops.	Item 1(ii) Specify the intended use of the AI intervention	Item 1a,b (ii) State the intended use of the AI intervention within the trial in the title and abstract
Describe the intended task of the AI intervention and its interaction with other healthcare professionals	8.0 (7.0-9.0)	7.0 (6.0-8.0)	100	0	100	0	These two items were merged and voted upon as one	Rewording issue. AI-human interface. This item overlaps with the next item and should actually be a subitem. What is the exact role of the AI intervention? What is it compared to? Specify this in the Explanation & Elaboration paper. Include public as well as healthcare professionals as intended users.	Item 6a (i) Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare professionals, patients, public).	Item 2a (i) Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare professionals, patients, public).
State the intended use of AI intervention in the context of the clinical pathway	8.0 (7.0-9.0)	8.0 (7.0-9.0)						It is important for this point to be accessible by the public; therefore, it should be included in the abstract.		
Describe prior (level) evidence for validation of the AI intervention	7.0 (6.0-8.0)	7.0 (6.0-8.0)	90	10	77	23	For CONSORT, not unique to AI interventions	This item is more suitable for SPIRIT than CONSORT (safety issues). Return to prior level of validation or feasibility/level of evidence and provide context for level of evidence. It should be clear if prior validation was for the same use/purpose.	Item 6a (ii) Describe any pre-existing evidence for the AI intervention	
Description of the onsite requirements needed to integrate the AI intervention into the trial setting and differences between trials sites	7.0 (6.0-8.0)	6.5 (5.0-8.0)	81	19	83	17		Often you don't know what the implementation process will be. Vital for SPIRIT, but not for CONSORT. Reporting of limitations of the model/cloud-based requirements is vital. Minimal requirements is useful to know. From a regulator's perspective, feasibility of implementation is important to know. This item is only relevant if outcomes are key to the infrastructure. There may be major limitations from localisation and replication challenges.	Item 9 Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting.	Item 4b Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements.
Describe the inclusion and exclusion criteria at the level of participants; and at the level of the input data	9.0 (8.0-9.0)	9.0 (8.0-9.0)	100	0	100	0	Proposed to split into two items (1. participants level 2. input data level) and voted upon both individually	Who and how are vital elements. Example: excluding participants on the basis of imaging quality. Is the quality enough for the algorithm (effectiveness). Input vs participants is not the same thing (i.e. cases vs imaging). It is important to clarify the inclusion and exclusion criteria of the study in order to increase authors understanding. This phase happens before randomization.	Item 10 (i) State the inclusion and exclusion criteria at the level of participants.	Item 4a (i) State the inclusion and exclusion criteria at the level of participants.
State which version of the AI algorithm is used, if relevant	8.0 (7.0-9.0)	8.0 (7.0-9.0)	90	10	93	7		Important to include the architecture of a deep learning model or include reference of a paper in which details of the algorithm are stated. Version of the AI algorithm is used to compare AI versions over time. This item will need revisiting soon. Include reference to regulatory papers. This item is essential from the regulatory perspective.	Item 10 (ii) State the inclusion and exclusion criteria at the level of input data.	Item 4 (ii) State the inclusion and exclusion criteria at the level of input data.
Indicate whether the trial setting is the same as the AI intervention development setting	7.0 (6.0-8.0)	7.0 (6.0-8.0)	30	70	26	74	Covered by SPIRIT item 9 and CONSORT item 4b	Any differences in methodology may be important, not exclusively the setting. Difference in performance across sites is common, probably already covered by current guideline but this is important enough for AI that it should be covered again. This item is not specific enough to be relevant. Already covered by CONSORT.		
Describe any interim analyses performed and any changes to the AI intervention	7.0 (7.0-9.0)	7.0 (6.0-9.0)	43	57	53	47	Covered by SPIRIT item 21b and CONSORT item 7b	Useful when you want to adapt the artificial intelligence model within the trial.		
Describe the rationales and assumptions for the sample size calculation	7.0 (6.3-9.0)	7.0 (5.0-9.0)	20	80	16	84	Covered by SPIRIT item 14 and CONSORT item 7a	Core CONSORT/SPIRIT guidelines may cover this already, depending on the trial. It is important to clarify in the Explanation & Elaboration paper.	Item 11a (i) State which version of the AI algorithm will be used.	
Specify sample size calculations carried out to determine reliable control arm intervention			35	65	29	71	Covered by SPIRIT item 14 and CONSORT item 7a	Sample size calculation may be different in artificial intelligence studies (i.e. variability across experts, where experts are the control intervention). This level of variability can have a significant impact on diagnostic validity. Experience of the diagnostician (as the control arm) makes a massive difference to the performance but is this really applicable to an RCT? Important point to include in the elaboration.		
Describe any patient involvement in trial design			58	42	48	52	Beyond scope.	This is not in the original CONSORT and SPIRIT guidelines. Public perception/public awareness is highly stressed, specially in funding applications. This is generic and not AI specific.		

Supplementary Table 2. Consensus meeting notes and decisions for SPIRIT-AI and CONSORT-AI

	SPIRIT-AI	CONSORT-AI	SPIRIT-AI		CONSORT-AI		SPIRIT-AI	CONSORT-AI	Reasons for exclusion	Consensus meeting discussion notes	Final SPIRIT-AI item	Final CONSORT-AI item
Candidate items arising from Delphi Surveys	Delphi median score (IQR)	CONSORT-AI Delphi median score (IQR)	(%) INCLUDE	(%) EXCLUDE	(%) INCLUDE	(%) EXCLUDE	Extension/Elaboration SPIRIT 2013	Extension/Elaboration CONSORT 2010				
Specify any planned ancillary analyses for subgroups where the algorithm is expected to show impaired performance			39	61	35	65		Covered by SPIRIT item 20b and CONSORT item 18	Covered by SPIRIT item 18	Stratification and subgroup analyses (biases i.e. ethnicity break down) are vital for AI. There are examples of papers which faced criticism because they did not provide substratification for ethnicity. Exploratory analysis is not new for CONSORT or SPIRIT. SPIRIT and CONSORT may cover this already.		
Specify the protocol for acquiring the input data for the AI intervention	8.0 (5.0-9.0)	7.0 (6.0-9.0)	83	17	84	16	Extension	Extension		How much data cleaning/pre-processing has been done? This is already covered by SPIRIT and CONSORT but it is essential to include it. From the regulatory perspective, this is important for auditing. The version of software should be reported; sufficient information is currently rarely provided. This information is always lacking during the peer review process. This allow you to judge any potential bias and important for replication. Always has to be requested.	Item 11a (iii) Specify the procedure for acquiring and selecting the input data for the AI intervention.	Item 5 (ii) Describe how the input data were acquired and selected for the AI intervention.
Specify the protocol in the case of missing input data	7.5 (6.0-9.0)	7.0 (7.0-9.0)	73	27	77	23		Revoted upon. Results below.		SPIRIT already covers this item. There are two types of missing data: what goes into the algorithm versus the outcomes in the trial that haven't been collected. This item refers to the part that goes into the algorithm. Pre-specifying the minimal level of quality is vital for safety. Withholding a prediction in case of missing data is also an option. This could be mistakenly confused with the eligibility criteria. Should this refer to the TIDIER checklist? To facilitate reporting, it should be reported in SPIRIT-AI and CONSORT-AI. This is very important as it is one major factor which can be controlled and is the advantage of doing a randomised controlled trial.	Item 11a (iii) Specify the procedure for assessing and handling poor quality or unavailable input data.	Item 5 (iii) Describe how poor quality or unavailable input data were assessed and handled.
Specify the protocol in the case of missing input data	7.5 (6.0-9.0)	7.0 (7.0-9.0)	97	3	97	3	Extension	Extension				
Specify the protocol for human-AI interaction	7.0 (7.0-9.0)	7.0 (6.0-8.0)								Only applies if there is an interaction. Specific to the role of the AI intervention. AI-human interaction is critical for ethics committees. If not defined, there may be down-stream confusion and problems understanding risks. It needs to be clear how management decision was arrived at. This scenario is similar in the case of a genetic test result: managing the consequences of a decision. Interface on the input and output sides are important. Instructions are very important on what to do with a test result. It is recommended to stick to the term "human", as it is broad enough. Helpful to assess risk/benefit ratio for regulators and the public. Specify how the artificial intelligence output is being used to influence decision-making. Patient non-compliance to recommendations also need to be captured.	Item 11a (iv) Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required of users.	Item 5 (iv) Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users.
Detail the required level of expertise of health care professionals and operators for interacting with the AI intervention	6.0 (5.0-7.5)	6.0 (5.0-7.0)	100	0	97	3	Extension	Extension	These two items were merged and voted upon as one.			
Specify what is the output of the AI intervention	9.0 (8.0-9.0)	9.0 (7.0-9.0)	100	0	100	0	Extension	Extension		There was no discussion around this item. Consensus participants decided to go straight to voting because of the high importance of the item.	Item 11a (v) Specify the output of the AI intervention.	Item 5 (v) Specify the output of the AI intervention.
Explain the protocol for how the AI intervention will lead to treatment decision-making	8.0 (6.0-9.0)	8.0 (7.0-9.0)	94	6	97	3	Extension	Extension		Actual regulatory decision. Treatment decision may not be done by the person that has interacted with the artificial intelligence intervention. Take out 'treatment' and make the decision making element more broad.	Item 11a (vi) Explain the procedure for how the AI intervention's outputs will contribute to decision-making or other elements of clinical practice.	Item 5 (vi) Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.
Describe the control intervention sufficient to allow replication			20	80	20	80		Covered by SPIRIT item 11a and CONSORT item 5	Covered by SPIRIT item 11a and CONSORT item 5	Control arms tend to be poorly reported (often just describer as "usual care"). Mandating for what is reported in the interventional arm should be reported in the control arm. It may be very expensive to get all this information (not necessarily mandatory). This item is already covered by SPIRIT and CONSORT, captured in description for intervention. .		
Provide an explanation of how uncertainty from the intervention will be communicated to end users			17	83	10	90		Beyond scope	Beyond scope	Explainability can lead to uncertainty and introduce bias. First read it as a technical requirement. Uncertainty: concerns related to communication can introduce bias in certain ways.		
State whether the AI algorithm is a static model, or if it is continuously evolving. If the latter, provide details	9.0 (7.0-9.0)	9.0 (7.0-9.0)	30	70	27	73		These two items were merged and voted upon together. Beyond scope.	These two items were merged and voted upon together. Beyond scope.	This is an evolving field and we are lacking tangible examples of continuously updating medical AI algorithms to provide guidance on this item. Technology is not ready yet, therefore we cannot provide any guidance. There's a huge danger of overfitting. It is important to have an item where we specify the algorithm version/ architecture/ evolving vs static (to be included in Explanation & Elaboration). Important to revisit SPIRIT-AI and CONSORT-AI in a few years.		
Describe the nature of continuous updating of the AI intervention; if relevant	9.0 (7.0-9.0)	9.0 (7.0-9.0)										
Describe the type of model and/or reference details of the AI algorithm			77	23	74	26		Beyond scope	Beyond scope	(New item generated during consensus meeting discussion and voted upon)		
In the case of continuously updating algorithms, describe the new training data								Beyond scope	Beyond scope			

Supplementary Table 2. Consensus meeting notes and decisions for SPIRIT-AI and CONSORT-AI

	SPIRIT-AI	CONSORT-AI	SPIRIT-AI		CONSORT-AI		SPIRIT-AI	CONSORT-AI	CONSORT-AI	Reasons for exclusion	Consensus meeting discussion notes	Final SPIRIT-AI item	Final CONSORT-AI item
Candidate items arising from Delphi Surveys	Delphi median score (IQR)	CONSORT-AI Delphi median score (IQR)	(%) INCLUDE	(%) EXCLUDE	(%) INCLUDE	(%) EXCLUDE	Extension/Elaboration SPIRIT 2013	Extension/Elaboration CONSORT 2010					
In the case of continuously updating algorithms, report the level at which the data was partitioned for training and for validation/testing									Beyond scope	Beyond scope	Beyond scope		
State any deviations from trial protocol			NA	NA	52	48			Not unique to AI interventions		It is good to be transparent about the deviations. This is currently not captured but will be when the SPIRIT-AI and CONSORT-AI are revised in the future. Add examples. Link to regulatory guidance in jurisdiction.		
Report instances of misuse of the AI intervention recommendations, if relevant	7.0 (5.0-7.0)	7.0 (6.0-9.0)	45	55	47	53			Beyond scope		How would people report that? Analogous to cross-over/ intervention that wasn't used in the way it was intended. Misused against intended use. From the regulatory perspective, it is important to state the reason why the AI intervention was misused. In a report this will be important (incidences/adverse events). This item is already covered by SPIRIT. This is not specific to artificial intelligence. It is important to know why something wasn't adhered to.		
Describe the procedures and any occurrences of data breach	7.0 (6.0-9.0)	7.0 (6.0-9.0)	32	68	26	74			Covered by SPIRIT item 22 and CONSORT item 19		This item does not seem to differ from SPIRIT or CONSORT. SPIRIT: procedures in the event of any data breach. CONSORT: any occurrences of data breach.		
Where the AI intervention is a diagnostic or predictive model, provide a detailed summary of the false positives and false negatives			87	13	90	10	Extension	Extension	These two items were merged and voted upon together		Error analysis is vital (i.e. stratification due to ethnicity). This applies in the case of re-training due to systematic error (accuracy as part of the trial). Posthoc analysis is vital - people behaving unpredictably in each arm can be scrutinised. Identify subgroups in which the artificial intelligence should not be deployed in order to identify all errors and risk mitigation strategy vital.	Item 22 Specify any plans to identify and analyse performance errors. If there are no plans for this, justify why not.	Item 19 Describe results of any performance errors and how were identified, where applicable. If no such analysis was planned or done, justify why not.
Describe anticipated undesirable outcomes and risks, including worst-case scenario													
Use of AI should be explicitly described in consent materials	8.0 (6.0-9.0)	7.0 (5.0-9.0)	27	73	21	79			Beyond scope		This item is not unique to artificial intelligence. Ethics panel should decide whether artificial intelligence should be explicitly described in the participant consent form.		
State whether participant data can be safely withdrawn from the clinical trial, if needed			3	97	NA	NA			Beyond scope		Data can not be fully withdrawn and should be mentioned in the participant consent form. This item is not unique to artificial intelligence.		
Interpret results in the context of differences between the dataset used to develop and validate the AI intervention and the clinical trial data	NA	7.0 (6.0-9.0)			21	79			Covered CONSORT item 21		Artificial intelligence is specific in the sense that the intervention can be improved with every intervention. However, CONSORT already covers this item. Not necessarily unique to AI. Provide minimum list of things to report and examples of types of biases.		
Explain the underlying assumptions and mechanisms of the AI intervention and uncertainties of the results	NA	8.0 (6.0-9.0)			19	81			Covered by CONSORT item 20		Mandate some a priori analyses. Combine generalisability/bias analyses: input data, population and setting. This point is not about generalisability, which would happen in the future. This point is about pre-validation. Authors will likely explain under performance. This item is unique and it complements the point on versioning of the algorithm.		
Describe potential biases stemming from the included participants/data	NA	8.0 (7.0-9.0)			48	52			Covered by CONSORT item 20		Regulators want to know what devices/software were used. Important to include this in the Explanation & Elaboration paper. Minimum list of things that should be reported and examples of types of biases.		
If applicable, plans for any attempts to audit, decode or explain the AI intervention's recommendations	6.0 (6.0-9.0)	6.0 (5.0-9.0)	60	40	47	53			Covered by SPIRIT item 20b CONSORT item 22		Important to identify biases of the dataset. Interpretability may be harmful in certain cases. Currently explainability methods are not understandable in a straight forward way, however this is an issue that is unique to AI. Prespecification is vital. This should be done at an earlier stage - i.e. before the clinical trials stage. Explainability can inappropriately confer trust. Some situations where explainability is more tractable. Authors should state they will do it, but unreasonable to ask for prespecified analysis. Not to be seen as endorsing something that is unclear.		
Availability of the AI Intervention Code	7.0 (5.0-9.0)	7.0 (5.0-9.0)	100	0	100	0	Extension	Extension			It is important to release the architecture code and parameters for transparency purposes. Data sharing is useless without the coding. Funders perspective: it is important to share the code, specially if doesn't so it can be used/replicated. Availability of the the coding doesn't mean the AI model would be easy to replicate. It should be stated if the coding is available and under what license. Important to mandate commercially availability, which regulator approved it, unique identifier and which class. Not unique to AI. This item is not advocating for code sharing, but rather just to declare whether code is available.	Item 29 State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.	Item 25 State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.
Patents and patent applications for the AI intervention	6.0 (6.0-9.0)	6.0 (5.0-7.0)	7	93	7	93			Beyond scope				

Supplementary Table 2. Consensus meeting notes and decisions for SPIRIT-AI and CONSORT-AI

Candidate items arising from Delphi Surveys	SPIRIT-AI Delphi median score (IQR)	CONSORT-AI Delphi median score (IQR)	SPIRIT-AI		CONSORT-AI		SPIRIT-AI Extension/ Elaboration SPIRIT 2013	CONSORT-AI Extension/ Elaboration CONSORT 2010	Reasons for exclusion	Consensus meeting discussion notes	Final SPIRIT-AI item	Final CONSORT-AI item
			(%) INCLUDE	(%) EXCLUDE	(%) INCLUDE	(%) EXCLUDE						
Role of the AI developer	6.0 (6.0-6.0)	6.0 (5.0-7.0)	3	97	0	100			Covered by SPIRIT item 28 CONSORT item 25	This item already covered by SPIRIT - authorship section. In addition, the item is not unique to artificial intelligence.		
Describe the role of the sponsor			0	100	0	100			Covered by SPIRIT item 28 CONSORT item 25	It is not an artificial intelligence specific item. It is already covered by existing guidance.		