EXPLANATION AND ELABORATION (E&E)

Table of Contents

AI-SPECIFIC ITEMS 2

Title and abstract - Item 1 (Title) 2
Introduction - Item 2a-b (Intended use) 3
Methods - Item 3a-c (Participants) 4
Methods - Item 4a-c (AI system) 5
Methods - Item 5a-b (Implementation) 7
Methods - Item 6a-b (Safety and errors) 9
Methods - Item 7 (Human factors) 10
Methods - Item 8 (Ethics) 11
Results - Item 9a-b (Participants) 11
Results - Item 10a-b (Implementation) 12
Results - Item 11 (Modifications) 13
Results - Item 12 (Human-computer agreement) 14
Results - Item 13a-b (Safety and errors) 15
Results - Item 14a-b (Human factors) 16
Discussion - Item 15 (Support for intended use) 17
Discussion - Item 16 (Safety and errors) 17
Statements - Item 17 (Data availability) 18

GENERIC ITEMS 19

Title and abstract - Item I (Abstract) 19
Introduction - Item II (Objectives) 19
Methods - Item III (Research governance) 19
Methods - Item IV (Outcomes) 20
Methods - Item V (Analysis) 20
Methods - Item VI (Patient Involvement) 21
Results - Item VII (Main results) 21
Results - Item VIII (Subgroups analysis) 21
Discussion - Item IX (Strengths and limitations) 22
Statements - Item X (Conflicts of interest) 23

REFERENCES 24

AI-SPECIFIC ITEMS

Title and abstract - Item 1 (Title)

Identify the study as early clinical evaluation of a decision support system based on artificial intelligence or machine learning, specifying the problem addressed.

The intention of this item is to enable efficient identification and retrieval of the study during literature searches. Fundamental to this are statements in the title regarding: (i) the use of machine learning (ML)/artificial intelligence (AI) in the decision support system; (ii) the problem the decision support system is addressing, and (iii) the study stage. There are not, as yet, widely accepted definitions for the stages of AI studies (compared to, for example, "phase I" or "phase II" for drug development). While the DECIDE-AI guideline does not advocate any specific nomenclature, and in the anticipation of more standardised stage names, consistent use of the same terms describing initial clinical investigations would simplify study identification, and might include:

- "early-stage clinical evaluation": generic term referring to the position of initial clinical investigations at the beginning of the clinical evaluation pathway.
- "formative clinical evaluation": this describes a clinical evaluation process during which monitoring and feedback are collected in order to improve performance. It stands in contrast to the final, summative (and typically larger scale) evaluation process.
- "phase 2": Park et al. proposed an AI development pathway similar to other medical innovations, in which phase 2 studies have as objectives the "controlled AI system performance/efficacy evaluation by intended users in medical setting" as well as the improvement of the interface design and system quality[1].
- "stage 2": The IDEAL-D framework for medical devices evaluation advocates for a stage 2 (merging the original IDEAL stages 2a and 2b), whose main goal is to facilitate "progression to definitive randomised controlled trials"[2].

Relating to point (i), authors should include, in either the title or the abstract, both the name of the underlying algorithm, and the commercial name of the AI system where one has been assigned. The former is necessary for curating methodologically similar approaches, whereas absence of the latter will hamper efficient retrieval of all papers relating to a single commercial system. If limited by the title character count, part of this information could also be reported in the abstract. When the name of the underlying algorithm might not be familiar to a clinical audience (e.g. regularisation-based extreme gradient boosting algorithm), authors should consider using a more generic term in addition, like "machine learning" or "artificial intelligence". Information about the study design can also be added into the title depending on the target journal, or could instead be included in the abstract.

Introduction - Item 2a (Intended use)

Describe the targeted medical condition(s) and problem(s), including the current standard practice, and the intended patient population(s).

Items 2a and 2b describe the intended use (or intended purpose) of the AI system and related use specifications. This information relates to the intended use at scale and should not be confused with the actual use of the AI system during the study, which is described under the methods section. A clear description of the intended use is important for readers to contextualise the study and appraise whether the AI system use during the study is representative of the intended use. It is also useful for regulators, who sometimes refer to the intended use claimed in clinical studies to inform their decisions about the classification and approval of new devices. If the intended use of the clinical study is different from the intended use during preclinical development studies (see item 4a), this should be clearly stated. Definition of the target medical condition (e.g. sepsis) and associated clinical problem (e.g. finding the optimal balance between fluid and vasopressor dose) allow comparison against competing AI systems. Because standard practice (see glossary) may vary across different geographies, it is important to clearly describe the standard practice that the authors are comparing against. Providing clinical outcomes for the current standard practice in the evaluation environment could further support the appraisal of the AI system's potential added value in clinical practice. Definition of the target population is needed to interpret the relevance and generalisability of the study findings. Any known contraindications to the use of the AI system should also be reported under this item.

Introduction - Item 2b (Intended use)

Describe the intended users of the AI system, its planned integration in the care pathway, and the potential impact, including patient outcomes, it intends to achieve.

This item aims to provide information about the intended implementation of the AI system. The description of intended users (see glossary) should include any characteristics likely to influence their interaction with the AI system (e.g. user role and responsibilities in the healthcare system, specialty, level of training, familiarity with digital technology, or any specific expertise required). Details about the planned integration in the care pathway at scale might include the environment in which the AI system will be used, the ease of access to the AI system, the decision it will support, the level of human control, and the timing of the decision support. Reporting on the intended impact helps readers to appraise whether the study performance metrics were appropriately chosen. For an AI system aimed at improving patient care, authors should state which patient outcomes (e.g. 30-day hospital re-admission or mortality rate) are targeted.

Methods - Item 3a (Participants)

Describe how patients were recruited, stating the inclusion and exclusion criteria at both patient and data level, and how the number of recruited patients was decided.

Transparent reporting of the recruitment process is important for readers to appraise the risk of selection bias. Information about the recruitment strategy (active, passive, open access use in the community), sampling method (consecutive, random, etc.) and the procedure used to obtain consent (or its waiver, see item III of the generic item list) should be provided. In the field of clinical AI, patients are represented by the data they generate; this representation can be of different form and quality, potentially influencing patient inclusion in the study and/or the AI system outputs. For example, a participant could meet all inclusion criteria at patient level, but be excluded from the study due to low quality or incomplete data acquisition. Therefore, it is important to describe inclusion and exclusion criteria at both patient and data level, echoing the recommendations made by other AI reporting guidelines[3]. Data level criteria can include: acquisition time, acquisition technique, data quality, data completeness, and data format.

At early-stage, a formal statistical sample size calculation is not always required. However, a conscious decision is made on the sample size and should be explained transparently. Stating the initial recruitment target for example is helpful to identify any premature termination of the recruitment process (e.g. lack of interest from participants, issues related to funding, safety reasons). A general rationale for the number of patients recruited also helps to judge whether the size of the study population was adequate to answer the main research questions. This number might be based on comparable studies, logistical/time constraints, or a target number of patients with specific characteristics (e.g. metastatic cancer), and should consider how representative the included sample is, with respect to the intended patient population.

Methods - Item 3b (Participants)

Describe how users were recruited, stating the inclusion and exclusion criteria, and how the intended number of recruited users was decided.

AI systems are complex interventions, for which the interaction between the users and the AI system have an influence on the overall performance and the study results. An analogy can be drawn with surgical innovation, for which the IDEAL framework details the importance of considering both patient and operator characteristics[4,5]. Therefore, it is crucial to consider users as a defined study population in its own right. Similarly to item 3a, authors should describe the recruitment process in sufficient detail for the reader to appraise any selection bias. Because both patients and users are considered to be participants (see glossary), details about the procedure to obtain consent (or its waiver, see item III of the generic item list) should also be reported for users. However, unlike the previous item, exclusion at data level is generally not recommended because user data quality can be informative with regard to the AI system usability (difficulty in use, lack of interest, etc.; see items 7 and 14a).

There is no clear guidance on how best to select the number of users in early clinical investigations. However, in a similar manner to the participating patients, a conscious choice is made about the number and characteristics of user recruited. This choice should be clearly reported, accompanied by a rationale, and guided by how representative the included users are with respect to the intended end user population. Any involvement of the users in the AI system design (or participation of research team members as users) should also be transparently reported, because it could bias the study findings. The number of users in existing clinical AI studies is often low and frequently insufficient to support the claims made by the authors[6,7]. In some specific cases, patients may also be the users of the AI system.

Methods - Item 3c (Participants)

Describe steps taken to familiarise the users with the AI system, including any training provided prior to the study.

This item should enable readers to anticipate how much time and work will be required before the AI system can be used reliably and with confidence: learning curves are an important aspect of any new AI system (see item 14b). Training and familiarisation also play an important role in reducing differences in performance between users based on their previous exposure to similar technologies. Authors should report the type and number of training and practice sessions, including details of the cases presented during training, and the time allocated for sessions. Ideally, a training protocol and the training materials used should be made available as supplementary information or deposited in open science repositories, adapted if necessary to preserve developers' intellectual property. Occasionally, familiarisation may happen during a trial ('wash in') period. This usually entails users interacting with the AI system (with or without supervision) but with no active data collection. In such cases, the conditions of the trial period should be described in enough detail to allow replication. Familiarisation might also have already occurred through previous experience with the same or similar AI system and should then be described in similar term as a trial period.

Methods - Item 4a (AI system)

Briefly describe the AI system, specifying its version and the type of underlying algorithm used. Describe, or provide a direct reference to, the characteristics of the patient population on which the algorithm was trained and its performance in preclinical development/validation studies.

Authors should provide a concise description of the AI system. This includes naming the type of underlying algorithm (i.e. mathematical model) as well as describing the supporting hardware and, if relevant, supporting software. A full description of the mathematical model itself is not expected. Ideally, references to the development and validation studies should provide additional information about: the variable components of the algorithm (e.g. hyperparameters, kernels, neural network architecture), the environment and characteristics of the patient populations in the development/validation studies, the outcomes of interest and the performance metrics used in these

studies, and the study results. If this information was not previously published, it should be described in the manuscript or using an open science repository. This information is important to appraise the differences between the populations on which the AI system was developed/validated and the current study population, as well as any variation in performance. AI system facts labels (i.e. standardised descriptions of an AI system designed to facilitate the communication to its users of essential facts and inform the incorporation of the system recommendations into their decision-making) have been proposed in the literature[8] and offer a practical way to describe the relevant information concisely.

AI systems are typically developed through a number of sequential versions, so a clear identification of the version being evaluated (or versions, see item 12) is important to ensure comparison of results between different studies: identification may be by a system version number or, if available, a regulatory marker such as a unique device identifier. The description of the supporting hardware platform should allow the reader to understand the benefits and limitations of the platform in clinical settings (e.g. size, autonomy, ease of cleaning, etc).

If specific algorithmic thresholds are used in the study (e.g. cut-off set to not exceed a false positive rate of 15%), these should be described and a rationale provided. Such thresholds are indeed often imposed to reflect clinical requirements dependent on the implementation settings (see items 5a and 5b), and can have a direct impact on the AI system outputs and the study results.

Methods - Item 4b (AI system)

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

The information reported in this item is necessary for other researchers to evaluate the transferability of the findings to other settings. The extent of details required may depend on the context of the individual study and may include:

- a list of the data items, with units if relevant, used as input during the study (i.e. participants data used by the AI system to produce an individualised output)
- the timeframe of data acquisition. The acquisition period can provide important information in the context of iterative modifications and performance changes overtime (see item 11).
- the origin of the input data, for example, including whether the data were already routinely collected and present in the electronic health record (EHR) or whether new data needed to be collected
- a description of how specific data items were measured and/or of the acquisition settings (e.g. computed tomography scanner model and slices count). The protocols and devices used to acquire data can be important sources of variability and confounding (e.g. reference range, sensibility/sensitivity, resolution, data quality, additional features). Because AI models may base their learning on these differences[9,10], in addition to the actual underlying clinical data, it is essential to transparently report data acquisition characteristics.

- the method used to input the data into the AI system, for example, including whether data were automatically extracted from an EHR or required manual entry. This pertains to data input during the study - not for data input into the EHR itself or during algorithm training. Details about the way data were entered allows readers to appraise whether issues around potential input errors were adequately addressed.
- a description of the data pre-processing and how any missing values were defined and handled. Authors should consider issues around clinical application, scale-up, and algorithmic fairness when choosing the most appropriate way to define and handle missing data. For example, a lab variable can be considered missing if no data point is available within a window of 24 hours, or 72 hours, which have different implications for clinical practice. This should be either described fully, or reference made to previous studies, if the same procedures have been used.

Methods - Item 4c (AI system)

Describe the AI system outputs and how they were presented to the users (an image may be useful).

Authors should provide a clear description of the human-computer interface, or direct reference to development studies, to enable some appraisal of the output-specific user experience, including comprehensiveness, clarity and overall design. The way information is displayed plays an important role in how users will interact with the system[11] and might have regulatory consequences in the future, namely when deciding whether a CDSS should be considered as medical device[12]. Details should be given about the type and number of outputs of the AI system (e.g. the AI system segmented and gave a probability of malignancy for each detected nodule), and the design for any display interface: one or more images/screenshots/illustrations may be the most concise way of describing this. Additional information provided by the AI system, such as visualisation of the attention mechanisms, quantification of output uncertainty, or data allowing the contextualisation of the output (e.g. display of the variables most influencing the AI system recommendation) should also be described. Authors should also describe the level of customisation of the interface available to the users and any opportunity for them to provide interactive feedback to the AI system (for example about the relevance of the outputs or by manipulating the input data items to see the influence of including/omitting them on the AI system output).

Methods - Item 5a (Implementation)

Describe the settings in which the AI system was evaluated.

The environment in which the AI system was evaluated should be clearly described. This can include: the type and size of healthcare centre (e.g. major trauma centre), the location within the structure (e.g. emergency department), other relevant staff and technological support available (multidisciplinary trauma team, bedside radiography), or the physical availability of the AI system's supporting hardware (computer in the nursing office). Information reported under item 5a and 5b are important to

demonstrate whether the settings and conditions of the study are representative of the AI system intended use (see item 2a and 2b). When used in a live clinical environment, user decision-making will also be influenced by factors and information external to the AI system itself (e.g. verbal handover from other clinical staff or results of related diagnostic tests). A clear description of the evaluation settings is therefore important for the reader to fully appreciate this context.

Methods - Item 5b (Implementation)

Describe the clinical workflow/care pathway in which the AI system was evaluated, the timing of its use, how the final supported decision was reached and by whom.

Authors should provide information about how exactly the AI system was used during the study. Details about the integration in the clinical workflow/care pathway might include: the initial situation of patients and their reason for receiving care, the clinical decisions made using the AI system (including its significance: to treat or diagnose, to drive clinical management, or to inform clinical management)[13], any actions that users were instructed to take based on specific AI system outputs, the chronology of other relevant diagnostic tests or clinical decisions made along the care pathway, or the different diagnostic/therapeutic options available to address the problem at hand. Because cognitive biases, like the anchoring bias, can influence the weighting of AI system recommendations depending on their timing in the decision-making process, a description of the timing of decision support (concurrent vs. second reading) should be reported. In concurrent reading mode, users see the AI system recommendations at the same time as they receive the other information on which their decision-making is based. In second reading mode users first make a decision about a case, then see the AI system recommendations, and re-evaluate their initial decision in the light of this new information.

Decision-making is often shared between different healthcare professionals and their patients. In many jurisdictions, autonomy entails respect for patients' decisions even when these are not medically optimal. As such, patients' choices may conflict with medical recommendations with and without AI. Issues around liability might also bias the final clinical decision and steer it toward legally safer options[14]. In both cases, the final clinical decision would be influenced by factors external to the AI system. Therefore, authors should describe the general decision-making process, including what parties were involved, at what stage, and who was responsible for the final clinical decision (i.e. the decision impacting patient care). For example, an AI system might be used to suggest therapeutic options or additional testing for the treatment of prostate cancer based on lab results and imaging. These cases may then be presented to a cancer multi-disciplinary meeting which may or may not recommend the suggested options; and the final therapeutic decisions may then be made between clinicians and patients in the clinic, which may include a choice to forego all treatment.

Methods - Item 6a (Safety and errors)

Provide a description of how significant errors/malfunctions were defined and identified.

Errors and malfunctions can be assigned to three main categories within which there may be overlap:

(i) algorithm errors (e.g. the recommendation describes a nodule as malignant when it is not)
(ii) malfunctions of the supporting software/hardware (e.g. failure to produce a recommendation at all due to data extraction issue or empty battery)
(iii) use errors, or in other words errors involving users (e.g. the user input the wrong patient details or applied the AI system outside the medical indication for use).

How these errors and malfunctions were defined, how their significance was determined, and the standards (if any) against which they were judged, is context-dependent and should therefore be described by the authors. Information about the methods used to identify errors and malfunctions is also important to appraise the reliability of their reporting. A clear definition of errors and malfunctions is necessary to incorporate them in the AI system safety evaluation (see item 6b).

Methods - Item 6b (Safety and errors)

Describe how any risks to patient safety or observed instances of harm were identified, analysed, and minimised.

Safety assessment is a critical part of any clinical evaluation and a continuous process which occurs before, during and after clinical studies. As such, AI system manufacturers are expected to have a risk management process in place and to have conducted a risk assessment before implementing a system in clinical settings. Despite the most robust pre-clinical risk assessment, evaluating an AI system in a live clinical environment may uncover new risks or harms that need to be identified and analysed; and which then require mitigation strategies (both during the study and before progressing to larger efficacy studies). Clear reporting of the methodology used to identify and analyse risks/harms is necessary to appraise the robustness of the safety evaluation. Current regulatory frameworks analyse risks in the context of[13,15]:

(i) the likelihood of an event to occur,
(ii) its potential impact on participants,
(iii) the extent to which it would be detectable
(iv) the severity of the targeted medical condition.

The International Society for Standardization (ISO) and the International Electrotechnical Commission (IEC) offer further guidance and recognised standards on: risk management for medical devices (ISO 14971:2019 and ISO/TR 14971:2020); software life cycle process requirements (IEC 62304:2006); and good clinical practices for the design, conduct, recording and reporting of clinical investigations (ISO 14155:2020)[16–19]. It should be noted that safety evaluation is an important part of the medical device regulatory approval process, which has specific requirements that fall outside the scope of DECIDE-AI.

Ideally, the authors should also reference the relevant findings of the preclinical risk assessment of the AI system (including anticipated adverse effects and other risks associated with participation) and derived safety requirements. This information is important to judge whether, during the study, the AI system complied with these pre-established safety requirements[20]. Risk mitigation strategies used during the study should also be reported.

Methods - Item 7 (Human factors)

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

As with safety, human factors (see glossary) evaluation should already have been performed at a preclinical stage. DECIDE-AI covers the continued evaluation appropriate to the use of the AI system under new, live clinical conditions. Safety-related aspects of human factors evaluation will be a core aspect of the medical device regulatory process, but broader considerations of human-computer interactions and system integration are crucial for effective implementation of AI systems. Understanding and optimising the conditions surrounding the use of an AI system can not only improve the system performance but also increase its acceptance by future users.

The most appropriate human factors evaluation is context and device dependent, and general guidance exists, especially around usability testing. Usability evaluations should use validated tools where available and not be limited to evaluation of user satisfaction alone. Relevant evaluation could include: time to task completion, workload analysis (e.g. using the NASA-TLX[21,22]), display interface or user satisfaction questionnaires. ISO and IEC standards, such as ISO 16982:2002, ISO 9241-11:2018, IEC 62366-1:2015 and IEC/TR 62366-2:2016, as well as British Standards Institution (BSI) standards, such as BS EN 62366-1:2015+A1:2020 offer a non-exhaustive list of recognised standards, frameworks, and/or guidance to select an appropriate approach[23–27]. Available human factors tools, methods and frameworks offer additional information and standardized terminology to perform and describe these investigations[28,29].

The methods described under this item should also convey how the results reported under items 12 (human-computer agreement) and 14b (learning curves) were produced. For overall agreement, numerous well establish methods exist (e.g. Kappa indexes) and the present guideline can be completed with specific items from the Guidelines for Reporting Reliability and Agreement Studies (GRRAS)[30]. For further elicitation of the reasons for users' reactions to the AI system recommendations (see item 12), a combination of qualitative and quantitative human factors methods are available and should be described by the authors.

Learning curves are evaluated using the chronological evolution of other metrics as proxies (e.g. the percentage of uses that conform to the implementation protocol, the time needed by a user to make a decision, the perceived workload of using the system, the agreement rate with the AI system recommendation, or the rate of correct decisions). Which proxy metrics are best suited to the learning curve evaluation is context-dependent and should be defined by the authors.

Beside the methodology used, it is also important to describe the use cases (i.e. examples of typical use of the device) evaluated and how the users involved in the human factors evaluation were selected. These should represent the most common types of use and users. Authors should report any consultation with or involvement of human factors specialists.

Methods - Item 8 (Ethics)

Describe whether specific methodologies were utilized toward an ethics-related goal (such as algorithmic fairness) and their rationale.

Ethics methodologies refer to a constellation of practices to detect, quantify, and mitigate bias in algorithm outputs including, but not limited to, algorithmic fairness (computational adjustments attempting to correct for bias[31]). Given that these methodologies can affect algorithm accuracy[32,33], it is essential to consider and report whether any were used when evaluating the AI system for accuracy and outcomes in a live clinical context[34]. Application of ethics methodologies provide important context for the evaluation of AI systems when comparing them against current practice during early clinical evaluation. For example, an algorithm intended to conduct risk assessments for cardiac surgery may be adjusted given that the reference standard systematically increases the estimated risk in black patients[35]. Applying algorithmic fairness, then, would have implications for comparisons against the status quo and for outcomes as a function of patient race[34]. The rationale behind the use of ethics methods should be described in relation to the intended goal. Authors should report any consultation with or involvement of ethicists, domain experts, or advisory groups.

Results - Item 9a (Participants)

Describe the baseline characteristics of the patients included in the study, and report on input data missingness.

A description of the study population is important, to allow a comparison with the algorithm training set population and the patient population of any subsequent large scale efficacy evaluation or implementation. Dataset shift (i.e. changes in the distribution of underlying data between an algorithm's training and test sets) is an important concern when moving from algorithm development to clinical evaluation, or between clinical settings. Small differences in the underlying datasets can result in significant variation in the AI system outputs, which have implications for both the algorithm performance and patient safety[36,37]. The choice of baseline characteristics to be reported should be informed by the intended use of the AI system, factors known to have an influence on the outcomes of interest, and protected characteristics[38]. These might include: age, sex, ethnicity, socioeconomic status, geographical location, prevalence of the targeted medical condition(s), classification/severity of the targeted medical condition(s), risk factors for the targeted medical condition(s), key predictors included in the algorithm, or other relevant medical conditions.

Data missingness can impact on model performance and have ethical implications, for example if more prevalent amongst protected patient groups or patients with lower health literacy[39–41]. Its extent can vary between the controlled *in silico* testing environment and live clinical settings. Therefore, authors should quantitatively report how much of the patient data used as input to the AI system (see item 4b) were missing during the study, ideally broken down by data item.

Results - Item 9b (Participants)

Describe the baseline characteristics of the users included in the study.

As explained in item 3b, the characteristics of the user population play an important role in the overall AI system performance. For example, the AI system's influence on decision-making might be associated with the users' level of clinical experience[6]. A clear description of the user characteristics is necessary to appraise any potential selection biases. Authors should consider reporting the users' medical specialty, level of training, clinical role/seniority, familiarity with the decision at hand (e.g. yearly case load if available), and their prior exposure to the decision support tool or similar technology. In studies with a small number of users, authors should carefully consider how best to maintain users' anonymity while reporting their baseline characteristics.

Results - Item 10a (Implementation)

Report on the user exposure to the AI system, on the number of instances the AI system was used, and on the users' adherence to the intended implementation.

Implementation science refers to these aspects as "implementation reach", "implementation dose" and "implementation fidelity". Authors should report the proportion of potential users who actually had exposure to the decision support tool, and how often the tool was used by the users who had access to it. If not separately reported as use errors, failures to adhere to the instructed use of the AI system (e.g. the medical indication for use, timing of use, or AI system function used) should also be reported, because such deviations can have an impact on the study results and are informative for future larger-scale implementations. When appropriate, a brief description of cases in which the AI system should have been used but was not (whether intentionally or not) may also be helpful because it might shed additional light on the reasons for non-adherence to the instructed use.

Results - Item 10b (Implementation)

Report any significant changes to the clinical workflow or care pathway caused by the AI system.

The adoption of a new technology can have a disruptive effect on existing clinical workflows and care pathways (see glossary). Workarounds and changes in behaviour are not uncommon following the introduction of new interventions in clinical settings. Reporting changes in clinical workflow or care pathway is important to identify potential cofounders and inform choices about the appropriate study design for larger scale evaluation[42].

In the context of this item, clinical workflow (i.e. what the clinicians do for their patients and when) and care pathways (i.e. what the patients experience during their contact with the healthcare system) should be differentiated and reported separately. For example, the use of an AI system can, through additional workload due to manual data input, increase the time taken by the clinician to complete a ward round, hence modifying the clinical workflow; but this may not necessarily affect the patients' experience or affect it in a different way, like for example by decreasing the time to referral for specialist care. The choice of which important changes to report should be guided by:

- the divergence from the anticipated integration in the care pathways described under item 2b (e.g. the AI system intended to reduce the use of inappropriate imaging, but ended up increasing the volume of specialist referrals as an unforeseen consequence)
- the potential risk to patient safety (e.g. patients were overall exposed to higher dose of ionising radiations)
- the potential impact on integration and acceptance of the AI system (e.g. users had to spend 50% more time on the discharge summary to retrieve the information from the AI system)
- the potential confounding effect on the chosen outcomes (e.g. every enrolled patient received additional laboratory tests in order to generate input data for the AI system).

Results - Item 11 (Modifications)

Report any changes made to the AI system during the study. Report the timing of these modifications, the rationale for each, and any changes in outcomes observed after each of them.

Provided that they are permitted by current regulation, obtained ethics approval and study protocol, changes to the algorithm (e.g. recalibration) or its supporting hardware platform (display interface improvement) during the study can be acceptable, especially when there is no attempt to make an overall summative conclusion about device effectiveness. Rapid evaluation-design cycles are encouraged in user-centred design theory[43,44], because they provide the means to tailor a product to its user needs. This may include changes to the algorithm (e.g. recalibration) or changes to its supporting hardware platform (e.g. display interface improvement) among others. Nonetheless, such changes and their impact on the main study outcomes (see item IV of the generic item list) should be carefully registered, versioned, and reported to understand the evolution of the AI system and to avoid the repetition of mistakes.

The IDEAL framework[4,5], for example, describes a whole stage (2a) of complex intervention evaluation dedicated to the reporting of early-stage changes made to operative procedures by the primary researchers and their impact on key metrics. The framework goes further with stage 2b, in which small modifications made by early adopters of new procedures (ideally in other centres) are reported. Only once a common, multicentric and stable description of the intervention has been agreed upon should the evaluation progress to larger trials. While it may not be realistic to report every software patch made during implementation, authors should carefully consider whether a modification could have an impact on the study outcomes and they should report any potentially influential changes. Changes occurring outside of the study (e.g. an update in the EHR architecture, decrease in the blood culture processing time) can also influence the AI system performance and should be reported if considered relevant.

DECIDE-AI focuses on changes made to adapt AI systems to their users or specific implementation settings as part of the early development and evaluation process. Continuous learning, self-learning, long-term maintenance or updates, and auditing are outside the scope of the present guideline.

Results - Item 12 (Human-computer agreement)

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

Decision support tools cover a spectrum from providing tailored information about a case, albeit without a clear recommendation for what to do, all the way through to direct recommendations that impact upon care if acted on by the decision maker. Reporting on the user agreement with the AI system becomes easier the closer the evaluated system lies to the latter end of this spectrum. Therefore, item 12 only applies to AI systems whose outputs can be considered as a recommendation (e.g. benign/malignant classification), or when a threshold for decision-making was attributed (e.g. if likelihood of deterioration >= 0.9, then escalate care).

Analysing how users react to AI system recommendations, why they do so, and in which specific instances, are important for understanding the underlying mechanisms of the human-computer interaction. Investigation of these dynamics is key to appraise the intrinsic value of the AI system and the role played by human users. It can also help to improve trust and usability, and thereby the overall acceptance and effectiveness of the AI system. Quantifying how much users rely on the AI system recommendations to make their decisions might also play a role in regulatory decisions in the future[12].

Decision support systems' recommendations are designed to influence user decision-making. Depending on the reaction of the user to the AI system recommendations, three broad situations might occur: (i) no change in decision made/action taken, (ii) an improvement in decision made/action taken, in which case the potential added value of the AI system is highlighted, or (iii) a worsening in decision made/action taken, in which case the use of the AI system exposes the patient to additional risks.

However, limiting the reporting to the changes in decision made/action taken will obscure important information such as hidden additional risks created by the use of the AI system but mitigated by users and missed opportunities to improve care. E&E Table 1 describes some of these simplified scenarios, as examples.

Authors should report on the overall user agreement with the AI system, as well as on instances of and reasons for disagreement with the AI system and of users changing their mind based on the AI system recommendations. Important information may include: initial user decision, AI system recommendation, final user decision, clinical situation, patient/case characteristics, user characteristics, reasons for variation/changing mind, consequences of the variation/mind change.

Scenario	User opinion*	AI system output*	User behaviour	Result	Interpretation
1	Correct	Correct	Follows own opinion and AI recommendation	No change in decision made/action taken	Status quo
2	Correct	Incorrect/ worse than user	Diverges from AI recommendation	No change in decision made/action taken	Additional risk: mitigated by user§
3	Correct	Incorrect/ worse than user	Follows AI recommendation	Worsening in decision made/action taken	Additional risk: potential harm to patient§
4	Incorrect	Correct/ better than user	Follows AI recommendation	Improvement in decision made/action taken	Added value of AI system§
5	Incorrect	Correct/ better than user	Diverges from AI recommendation	No change in decision made/action taken	Missed opportunity to improve care§
6	Incorrect	Incorrect	Follows own opinion and AI recommendation	No change in decision made/action taken	Status quo

E&E Table 1. Simplified scenarios of user reactions to AI recommendations and their influence on the decision made/action taken. * Not all user opinions or AI system recommendations can be easily classified along an axis from correct/optimal to incorrect/suboptimal. § assuming causality between the decision made/action taken and clinical outcome of interest (e.g. a one-off change in fluid prescription may not impact on intensive care unit length of stay).

Results - Item 13a (Safety and errors)

List any significant errors/malfunctions related to: AI system recommendations, supporting software/hardware, or users. Include details of: (i) rate of occurrence, (ii) apparent causes, (iii) whether they could be corrected, and (iv) any significant potential impacts on patient care.

All observed types of significant errors/malfunctions, as described under item 6a, should be reported (ideally in a table) and briefly described, specifying:

(i) the number of occurrences observed
(ii) their apparent causes (a full audit of each error/malfunction is not within the scope of DECIDE-AI)
(iii) if they were detected before having a negative impact and any action taken to correct them
(iv) what impact they had, or would have had, on patient care.

During early clinical evaluation, it is expected that the number of significant error types remains low enough for a detailed failure case analysis. Numerous errors would strongly suggest the need to further refine the AI system, prior to repeat live evaluation. Researchers may also wish to breakdown the error/malfunction reporting according to relevant patient subgroups[38]. Transparent reporting and analysis of errors/malfunctions can allow other research teams to avoid encountering similar problems and thereby safeguard patients. It also informs product design improvement prior to subsequent large-scale trials.

Results - Item 13b (Safety and errors)

Report on any risks to patient safety or observed instances of harm (including indirect harm) identified during the study.

Authors should report on the identified risks to patient safety and observed instances of harm, according to the methodology described under item 6b. Risks and harms can derive not only from errors/malfunction or misuse of the AI system, but also from its intended and correct use. As in the early clinical phases of drug development, instances of harm include both expected and unexpected adverse reactions, both direct and indirect. Knowledge about such risks and instances of harm are crucial to appraise the safety profile of the AI system and to develop mitigation strategies before larger scale evaluation.

Results - Item 14a (Human factors)

Report on the usability evaluation, according to recognised standards or frameworks.

Usability is described by the International Society for Standardization as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"[24]. The most appropriate usability evaluation approach will be context-dependent, and should be guided by recognised standards or frameworks, using appropriate methodology (see item 7). The reporting of the human factors results should be guided by best practice in the chosen methodology. The characteristics of the participants in the human factors evaluation should be specified if different (or a subset) from the user population. Some aspects of usability testing will relate to safety and might therefore overlap with item 13b. If relevant, comparison against preclinical usability evaluation can also be described.

Results - Item 14b (Human factors)

Report on the user learning curves evaluation.

The existence of learning curves (see glossary) has been observed in the field of computer-assisted decision-making[45,46] and well documented for other medical interventions, especially in surgery[47,48]. Information about learning curves is crucial for the design of subsequent trials. Indeed, if a research team starts to collect data before users have reached a stable performance level with a new device, they are likely to bias their findings, often to the detriment of the evaluated intervention. Authors should report on the learning curves evaluation according to the methodology and metrics described under item 7. Summary statistics are helpful to understand the practical implication of the learning curves (e.g. a median of 94 cases – interquartile range: 85 to 108 – was necessary for users to reach a stable daily agreement rate with the AI system). A graphical representation of the learning curves can provide more granular information for reader trying to apply the study findings to different context.

Discussion - Item 15 (Support for intended use)

Discuss whether the results obtained support the intended use of the AI system in clinical settings.

Authors should describe what can realistically be expected of the evaluated system in light of the reported results, and how these results support the device's intended use, in comparison with the current standard practice (ideally providing some numeric benchmark performance metrics) and similar studies. Claims should be proportionate to the strength of the evidence generated, considering the study limitations (see item IX of the generic item list), and should avoid premature conclusions about the potential benefits of adoption[7]. Authors should discuss the key clinical performance findings in the context of the human factors evaluation results. Authors should also address the potential challenges in adoption of the AI system for larger comparative trials, and they should identify what questions remain to be answered, or product improvements made, while progressing to the next stage of evaluation. At this stage, justified modifications to the AI system indications for use, or even intended use, can be discussed and can inform future trials or the ongoing regulatory approval process[4,5].

Discussion - Item 16 (Safety and errors)

Discuss what the results indicate about the safety profile of the AI system. Discuss any observed errors/malfunctions and instances of harm, their implications for patient care and whether/how they can be mitigated.

Authors should summarise the key safety-related findings of the study, considering errors/malfunctions, identified risks, observed adverse events, unexpected changes in care pathways, and the results of safety-related human factors evaluation. Errors are to be expected during early-stage clinical evaluations, and their recognition and analysis is vital prior to larger trials. Possible mitigation

strategies in the context of future evaluation can be discussed, if relevant. These strategies can include algorithm retraining, further product development, or modified study design for subsequent trials. If discussed, a rationale for the choice of specific mitigation strategies should be provided, considering the available options and likelihood of mitigating the risk in subsequent trials. For example, if it is possible to act on a higher level (e.g. redesigning the AI system to be more user friendly), this should be pursued before moving to lower levels of mitigation (e.g. additional training sessions for users as a workaround for an AI system that is not particularly user friendly). A comprehensive and transparent reporting of identified safety issues, as well as an open discussion about how to mitigate or avoid them creates a robust safety culture, which in turn will increase user and public trust in the technology.

Statements - Item 17 (Data availability)

Disclose if and how data and relevant code are available.

Individual patient data and code are key components to facilitating reproducible science. There are of course limitations to the circumstances wherein both can be shared openly and without restriction. This item is a prompt to emphasize the importance of appropriately communicating whether the authors offer the possibility for the community to replicate their findings and verify their code (including both the algorithm and relevant supporting software code), and if not, why. If data and code have been shared, the manuscript should describe what level of access can be expected (for example using reproducibility standards[49]) and how this can be practically obtained. In the context of this item, data refers to data collected during the study (participant and outcome data), rather than the data used to train the algorithm itself.

GENERIC ITEMS

Title and abstract - Item I (Abstract)

Provide a structured summary of the study. Consider including: intended use of the AI system, type of underlying algorithm, study setting, number of patients and users included, primary, secondary, safety and human factors outcomes measured, main results, conclusions.

The abstract is an important screening tool, both for the general reader and systematic reviewers looking to determine whether the study matches their inclusion/exclusion criteria. In addition to reporting the usual major elements of a study, key AI-related points that merit inclusion within the abstract are:

- the intended use for the AI system (targeted medical condition(s) and problem(s); can also describe the intended patient population, intended users, and intended use environment)
- the underlying algorithm type (often an important selection criterion for systematic reviews)
- the number of users involved (given the influence of user variability on the AI system performance)
- the key safety endpoints (AI system safety evaluation in live clinical settings is essential to gain trust of both patients and users, and a key feature of early-stage clinical studies prior to larger scale evaluation)
- the assessment of human factors (because the results from these are likely to have as much importance in early-stage studies than clinical outcomes).

Introduction - Item II (Objectives)

State the study objectives.

The study objectives operationalise the research question that the study was designed to answer. Clearly stated objectives help readers to appraise whether or not the study design was appropriate, whether the resulting data from the study have fulfilled the objectives, and therefore whether a move to the next step of evaluation is warranted.

Methods - Item III (Research governance)

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

Authors should provide a reference to any journal article or open repository entry where the study protocol has been published. The existence of a protocol and its publication prior to data collection and analysis lower the risk of bias due to selective analysis, selective reporting of outcomes, and/or outcome switching. The Declaration of Helsinki states that "every research study involving human subjects must be registered in a publicly accessible database before recruitment of the first subject"[50]. Most ethical

review boards will consider study registration as a condition for approval of clinical trials - even small ones. For the other types of study, registration is considered good research practice. Both protocol publication and study registration improve research transparency and allow assessment of bias (e.g. reporting bias, publication bias). Authors should report the study ethics approval number and are encouraged to state whether informed consent (and assent, where relevant - for example, in pediatrics) was obtained or whether the ethics review board granted a waiver of consent.

Methods - Item IV (Outcomes)

Specify the primary and secondary outcomes measured.

In conventional hypothesis testing studies, the distinction between primary and secondary outcomes is that the study is appropriately powered to detect a statistically significant difference between groups for the primary outcome. Clearly describing whether an outcome is primary or secondary helps readers to place the results in the appropriate context. For early-stage AI studies, a similar distinction can also be followed if the study is powered appropriately for the primary outcome (e.g. a process measure which might require a smaller sample). In non-hypothesis testing studies, the primary outcomes are instead the outcome that the authors believe are the most important to the overall objectives of the study. Clearly identifying the study outcomes also allows comparison between studies and, if the outcomes are chosen adequately, the observation of performance evolution over sequential trials.

Methods - Item V (Analysis)

Describe the statistical methods by which the primary and secondary outcomes were analysed, as well as any prespecified additional analyses, including subgroup analyses and their rationale.

Most early-stage studies will be small and underpowered for significance testing of clinical outcomes. Nonetheless, important aspects of future trials, such as the most appropriate outcomes, the expected effect size, optimal inclusion and exclusion criteria for the patient and user populations, the evolution of the users' learning curves, and the best decision support timing, can be derived from prospective observational cohort studies. For AI systems providing recommendations, the statistical methods should describe how agreement with human users (see item 12 of the AI-specific item list) was derived. In cases where the AI systems were subject to modification, the description of the statistical methods should clarify how these modifications were accounted for in the analyses (e.g. by use of repeated Bayesian analysis). If statistical significance is reported, authors should specify whether both patient and user variability was accounted for (e.g. using mixed effects models). Whereas some degree of explorative analysis can be useful in early-stage studies, the main subgroups of interest (for example, vulnerable patient populations, specific user experience levels, or medical condition subtypes) should be identified *a priori* and with appropriate rationale provided. Prespecifying subgroups of interest for analysis will help in the preparation of subsequent comparative trials[4,5] and build trust in the AI system, by addressing ethical concerns about algorithmic fairness at the implementation level[51].

Methods - Item VI (Patient Involvement)

State how patients were involved in any aspect of: the development of the research question, the study design, and the conduct of the study.

Patient and public involvement (PPI) has become a more important priority for funders, medical journals and end users of research in recent years. Published literature and official guidance suggest that patient involvement can lead to research quality improvement, improve participant recruitment and play an important role to strengthen the relationship between the scientific community and the public[52–54]. Stating the nature of patient involvement allows readers independently to assess the degree to which the research has been shaped by the concerns and values of patients, who are ultimately the final recipients for many of the stated benefits of AI-driven decision support systems. Amongst other benefits to the research team, early PPI can help to select outcomes that are valued by patients, understand how the information generated by the AI systems should be communicated to them, or how data collection could be organised to minimised disruption in the care pathway, thereby increasing acceptance of the intervention during subsequent larger trials. Authors should report the roles played by patients, and to what extent they were involved, in the development of the research question, study design and conduct of the study.

Results - Item VII (Main results)

Report on the prespecified outcomes, including outcomes for any comparison group if applicable.

The measurements of the prespecified primary and secondary outcomes should be reported, including measures of variability. For composite or derived outcomes (e.g. specificity of a given test), the underlying measurements (in this case: true positives, true negatives, false positives, false negatives) should also be reported as comprehensively as possible, with further detail included in appendices if necessary. Comparison groups are not always necessary during early stage clinical evaluation: their relevance and suitability will depend on the main study objectives.

Results - Item VIII (Subgroups analysis)

Report on the differences in the main outcomes according to the prespecified subgroups.

Subgroups can refer to patients (e.g. presence of a specific medical condition, stage of the condition, demographic group, biomarker value), users (e.g. specialty training, level of experience, level of adherence, level of agreement with the AI system), AI system characteristics (e.g. output presentation, cases of incorrect output, level of data missingness), or settings (e.g. in/outpatient, hospital site, time of the day). Understanding the differences in outcomes between the prespecified subgroups of interest is important, to assess any prior assumptions about variation in performance of the AI system (see Item V), and their implications for clinical practice. Because most AI systems are developed based on a digital

reflection of the current healthcare practice, they will, if not design and evaluated properly, only embed and perpetuate the, known and unknown, inequalities of the healthcare system. It is therefore extremely important to assess from an early stage how fair AI systems are between the different patient groups in practice. Understanding differences in outcomes between subgroups is also necessary to tailor the design of subsequent comparative trials, and they can inform further product improvement. For example, if the AI system demonstrates a consistent improvement in junior clinicians' performance, but only a marginal improvement in the performance of their senior colleagues, one could consider limiting the intended use of the AI system to junior clinicians.

Explorative subgroup analyses can also be informative in some cases: for example, if the AI system unexpectedly demonstrates better or worse performance with a patient subgroup (e.g. patients with a specific subtype of cancer), this could inform the design of subsequent evaluation. These explorative analyses should be clearly identified as such, and differentiated from prespecified subgroup analyses. In any case, future restrictions to the indications for use should always consider the nature of the excluded groups and provide appropriate justifications. There is a balance to strike between optimising personalisation/improvement in outcomes for subgroups and exacerbating existing inequities in the provision of healthcare, especially for disadvantaged groups. Investigating and transparently reporting the nature of, reasons for, and potential implication of discrepancies in performance between subgroups is important prior to settling on a final indication for use/target population[51].

Discussion - Item IX (Strengths and limitations)

Discuss the strengths and limitations of the study.

Authors should discuss the key aspects differentiating the study from other works already published and they should highlight the aspects of the study design considered particularly robust. Known limitations in the study design, chosen methodology, or the results obtained (e.g. limited follow up opportunities, no statistical power calculation, or lower than expected adherence to the instructed use) should also be transparently stated.

Some important elements specific to AI and decision support merit discussion (see Results section of the AI-specific list). These include whether it was feasible to deliver the intervention as intended, whether enough users were exposed to the AI system, potential biases introduced by user characteristics or selection process, and the impact of learning curves and human factor assessments. Authors are encouraged to discuss the potential impact of limitations on the study results, as well as any mitigating measures taken.

Statements - Item X (Conflicts of interest)

Disclose any relevant conflicts of interest, including the source of funding for the study, the role of funders, any other roles played by commercial companies, and personal conflicts of interest for each author.

Disclosure of conflicts of interest (CoI) has become a standard requirement for most peer-reviewed journals and is important for assessing any possible risk of bias. Financial incentives are significant in the clinical AI industry, therefore transparent reporting of CoI is crucial to gain public and peer trust in AI research, and to avoid any suspicion of vested interests. Beside the source of funding and role played by funders (e.g. involvement in defining the research question, the study design or in the outcome analysis), authors should also report the roles played by any other commercial entities, for example in logistic support, sharing of intellectual property, or other non-financial contributions, as well as a more direct involvement in the conduct of the study. The International Committee of Medical Journal Editors and most medical journals provide further guidance on appropriate reporting of CoI.

REFERENCES

1. Park, Y. *et al.* Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* 3, 326–331 (2020).
2. Sedrakyan, A. *et al.* IDEAL-D: A rational framework for evaluating and regulating the use of medical devices. *BMJ* 353, i2372 (2016).
3. Liu, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat. Med.* 26, 1364–1374 (2020).
4. McCulloch, P. *et al.* No surgical innovation without evaluation: the IDEAL recommendations. *Lancet* 374, 1105–1112 (2009).
5. Hirst, A. *et al.* No Surgical Innovation Without Evaluation: Evolution and Further Development of the IDEAL Framework and Recommendations. *Ann. Surg.* 269, 211–220 (2019).
6. Vasey, B. *et al.* Association of Clinician Diagnostic Performance with Machine Learning-Based Decision Support Systems: A Systematic Review. *JAMA Netw. Open* 4, (2021).
7. Nagendran, M. *et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368, m689 (2020).
8. Sendak, M. P., Gao, M., Brajer, N. & Balu, S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit. Med.* 3, 41 (2020).
9. Badgeley, M. A. *et al.* Deep learning predicts hip fracture using confounding patient and healthcare variables. *npj Digit. Med.* 2, 31 (2019).
10. Zech, J. R. *et al.* Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Med.* 15, e1002683 (2018).
11. Dudley, J. J. & Kristensson, P. O. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, (2018).
12. US Food and Drug Administration (FDA). *Clinical Decision Support Software - Draft Guidance for Industry and Food and Drug Administration Staff*. https://www.fda.gov/media/109618/download (2019).
13. IMDRF Software as Medical Device (SaMD) Working Group. *'Software as a Medical Device': Possible Framework for Risk Categorization and Corresponding Considerations*. (2014).
14. Price, W. N. 2nd, Gerke, S. & Cohen, I. G. Potential Liability for Physicians Using Artificial Intelligence. *JAMA* (2019) doi:10.1001/jama.2019.15064.
15. US Food and Drug Administration (FDA). *Oversight of Clinical Investigations - A Risk-Based Approach to Monitoring*. (2013).
16. International Organization for Standardization. *Medical devices - Application of risk management to medical devices (ISO 14971:2019)*. (2019).

17. International Organization for Standardization. Medical devices - Guidance on the application of ISO 14971 (ISO/TR 24971:2020). (2020).

18. International Electrotechnical Commission. Medical device software - Software life cycle processes (IEC 62304:2006). (2006).

19. International Organization for Standardization. *Clinical investigation of medical devices for human subjects - Good clinical practice (ISO 14155:2020)*. (2020).

20. Hawkins, R. *et al. Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)*. (2021).

21. Hart, S. G. & Staveland, L. E. Developmnent of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. in *Human Mental Workload* (eds. Hancock, P. A. & Meshkati, N.) (North Holland Press, 1988).

22. Hart, S. G. NASA-Task Load Index (NASA-TLX); 20 Years Later. in *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (2006).

23. International Organization for Standardization. *Ergonomics of human-system interaction - Usability methods supporting human-centred design (ISO/TR 16982:2002)*. (2002).

24. International Organization for Standardization. *Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts (ISO 9241-11:2018)*. (2018).

25. International Electrotechnical Commission. *Medical devices - Part 1: Application of usability engineering to medical devices (IEC 62366-1:2015)*. (2015).

26. International Electrotechnical Commission. Medical devices - Part 2: Guidance on the application of usability engineering to medical devices (IEC/TR 62366-2:2016). (2016).

27. British Standards Institution. *Medical devices. Application of usability engineering to medical devices (BS EN 62366-1:2015+A1:2020)*. (2020).

28. Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C. & Jenkins, D. P. *Human Factors Methods*. (Ashgate Publishing, 2005).

29. Sujan, M., Baber, C., Salmon, P., Pool, R. & Chozos, N. *Human Factors and Ergonomics in Healthcare AI*. (2021).

30. Kottner, J. *et al.* Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *J. Clin. Epidemiol.* 64, 96–106 (2011).

31. Mitchell, S., Potash, E., Barocas, S., D'Amour, A. & Lum, K. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annu. Rev. Stat. Its Appl.* 8, 141–163 (2021).

32. Pfohl, S. R., Foryciarz, A. & Shah, N. H. An empirical characterization of fair machine learning for clinical risk prediction. *J. Biomed. Inform.* 113, 103621 (2021).

33. Park, Y. *et al.* Comparison of Methods to Reduce Bias From Clinical Prediction Models of Postpartum Depression. *JAMA Netw. Open* 4, e213909–e213909 (2021).

34. McCradden, M. D., Joshi, S., Mazwi, M. & Anderson, J. A. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit. Heal.* 2, e221–e223 (2020).

35. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in Plain Sight — Reconsidering the Use of Race Correction in Clinical Algorithms. *N. Engl. J. Med.* 383, 874–882 (2020).

36. Finlayson, S. G. *et al.* The Clinician and Dataset Shift in Artificial Intelligence. *N. Engl. J. Med.* 385, 283–286 (2021).

37. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* 21, 345–352 (2020).

38. McCradden, M. D. *et al.* Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *J. Am. Med. Informatics Assoc.* 27, 2024–2027 (2020).

39. Marshall, A., Altman, D., Royston, P. & Holder, R. Comparison of techniques for handling missing covariate data within prognostic modelling studies: A simulation stud. *BMC Med. Res. Methodol.* 10, 7 (2010).

40. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.* 169, 866–872 (2018).

41. Gianfrancesco, M. A., Tamang, S., Yazdany, J. & Schmajuk, G. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern. Med.* 178, 1544–1547 (2018).

42. Wiens, J. *et al.* Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* 25, 1337–1340 (2019).

43. International Organization for Standardization. *Ergonomics of human-system interaction — Part 210: Human-centred design for interactive systems (ISO 9241-210:2019)*. (2019).

44. Norman, D. A. *User Centered System Design*. (CRC Press, 1986).

45. Hock, D. *et al.* Virtual Dissection CT Colonography: Evaluation of Learning Curves and Reading Times with and without Computer-aided Detection. *Radiology* 248, 860–868 (2008).

46. Rodríguez-Ruiz, A. *et al.* Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. *Radiology* 290, 305–314 (2019).

47. Harrysson, I. J. *et al.* Systematic review of learning curves for minimally invasive abdominal surgery: a review of the methodology of data collection, depiction of outcomes, and statistical analysis. *Ann. Surg.* 260, 37–45 (2014).

48. Hopper, A. N., Jamison, M. H. & Lewis, W. G. Learning curves in surgical practice. *Postgrad. Med. J.* 83, 777 LP – 779 (2007).

49. Heil, B. J. *et al.* Reproducibility standards for machine learning in the life sciences. *Nat. Methods* 18, 1132–1135 (2021).

50. World Medical Association. World Medical Association Declaration of Helsinki. Ethical principles for medical research involving human subjects. *Bull. World Health Organ.* 79, 373–374 (2001).

51. McCradden, M. D., Stephenson, E. A. & Anderson, J. A. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat. Med.* 26, 1325–1326 (2020).

52. Brett, J. *et al.* Mapping the impact of patient and public involvement on health and social care research: a systematic review. *Heal. Expect. an Int. J. public Particip. Heal. care Heal. policy* 17, 637–650 (2014).

53. Russell, J., Greenhalgh, T. & Taylor, M. *Patient and public involvement in NIHR research 2006-2019: policy intentions, progress and themes*. (2019).

54. Hayes, H., Buckland, S. & Tarpey, M. *Briefing notes for researchers: public involvement in NHS, public health and social care research*. (2012).