CONSENSUS STATEMENT
https://doi.org/10.1038/s41591-022-01772-9

nature medicine

Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI

Baptiste Vasey[1,2,3] ✉, Myura Nagendran[4], Bruce Campbell[5,6], David A. Clifton[2], Gary S. Collins[7], Spiros Denaxas[8,9,10,11], Alastair K. Denniston[12,13,14], Livia Faes[14], Bart Geerts[15], Mudathir Ibrahim[1,16], Xiaoxuan Liu[12,13], Bilal A. Mateen[8,17,18], Piyush Mathur[19], Melissa D. McCradden[20,21], Lauren Morgan[22], Johan Ordish[23], Campbell Rogers[24], Suchi Saria[25,26], Daniel S. W. Ting[27,28], Peter Watkinson[3,29], Wim Weber[30], Peter Wheatstone[31], Peter McCulloch[1] and the DECIDE-AI expert group*

A growing number of artificial intelligence (AI)-based clinical decision support systems are showing promising performance in preclinical, in silico evaluation, but few have yet demonstrated real benefit to patient care. Early-stage clinical evaluation is important to assess an AI system's actual clinical performance at small scale, ensure its safety, evaluate the human factors surrounding its use and pave the way to further large-scale trials. However, the reporting of these early studies remains inadequate. The present statement provides a multi-stakeholder, consensus-based reporting guideline for the Developmental and Exploratory Clinical Investigations of DEcision support systems driven by Artificial Intelligence (DECIDE-AI). We conducted a two-round, modified Delphi process to collect and analyze expert opinion on the reporting of early clinical evaluation of AI systems. Experts were recruited from 20 pre-defined stakeholder categories. The final composition and wording of the guideline was determined at a virtual consensus meeting. The checklist and the Explanation & Elaboration (E&E) sections were refined based on feedback from a qualitative evaluation process. In total, 123 experts participated in the first round of Delphi, 138 in the second round, 16 in the consensus meeting and 16 in the qualitative evaluation. The DECIDE-AI reporting guideline comprises 17 AI-specific reporting items (made of 28 subitems) and ten generic reporting items, with an E&E paragraph provided for each. Through consultation and consensus with a range of stakeholders, we developed a guideline comprising key items that should be reported in early-stage clinical studies of AI-based decision support systems in healthcare. By providing an actionable checklist of minimal reporting items, the DECIDE-AI guideline will facilitate the appraisal of these studies and replicability of their findings.

The prospect of improved clinical outcomes and more efficient health systems has fueled a rapid rise in the development and evaluation of AI systems over the last decade. Because most AI systems within healthcare are complex interventions designed as clinical decision support systems, rather than autonomous agents, the interactions among the AI systems, their users and the implementation environments are defining components of the AI interventions' overall potential effectiveness. Therefore, bringing AI systems from mathematical performance to clinical utility needs an adapted, stepwise implementation and evaluation pathway, addressing the complexity of this collaboration between two independent forms of intelligence, beyond measures of effectiveness alone[1].

[1]Nuffield Department of Surgical Sciences, University of Oxford, Oxford, UK. [2]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK. [3]Critical Care Research Group, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK. [4]UKRI Centre for Doctoral Training in AI for Healthcare, Imperial College London, London, UK. [5]University of Exeter Medical School, Exeter, UK. [6]Royal Devon and Exeter Hospital, Exeter, UK. [7]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, UK. [8]Institute of Health Informatics, University College London, London, UK. [9]British Heart Foundation Data Science Centre, London, UK. [10]Health Data Research UK, London, UK. [11]UCL Hospitals Biomedical Research Centre, London, UK. [12]University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [13]Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [14]Moorfields Eye Hospital NHS Foundation Trust, London, UK. [15]Healthplus.ai-R&D BV, Amsterdam, The Netherlands. [16]Department of Surgery, Maimonides Medical Center, Brooklyn, NY, USA. [17]The Wellcome Trust, London, UK. [18]The Alan Turing Institute, London, UK. [19]Department of General Anesthesiology, Anesthesiology Institute, Cleveland Clinic, Cleveland, OH, USA. [20]The Hospital for Sick Children, Toronto ON, Canada. [21]Dalla Lana School of Public Health, University of Toronto, Toronto ON, Canada. [22]Morgan Human Systems Ltd, Shrewsbury, UK. [23]Medicines and Healthcare products Regulatory Agency, London, UK. [24]HeartFlow Inc., Redwood City, CA, USA. [25]Departments of Computer Science, Statistics, and Health Policy, and Division of Informatics, Johns Hopkins University, Baltimore, MD, USA. [26]Bayesian Health, New York, NY, USA. [27]Singapore National Eye Center, Singapore Eye Research Institute, Singapore, Singapore. [28]Duke-NUS Medical School, National University of Singapore, Singapore, Singapore. [29]NIHR Biomedical Research Centre Oxford, Oxford University Hospitals NHS Trust, Oxford, UK. [30]The BMJ, London, UK. [31]School of Medicine, University of Leeds, Leeds, UK. *A list of members and their affiliations appears in the Supplementary Information. ✉e-mail: baptiste.vasey@gmail.com

Box 1 | Methodological challenges of the AI-based decision support system evaluation

The clinical evaluation of AI-based decision support systems presents several methodological challenges, all of which will likely be encountered at early stage. These are the needs to:

- account for the complex intervention nature of these systems and evaluate their integration within existing ecosystems
- account for user variability and the added biases occurring as a result
- consider two collaborating forms of intelligence (human and AI system) and, therefore, integrate human factors considerations as a core component
- consider both physical patients and their data representations
- account for the changing nature of the intervention (due to early prototyping, version updates or continuous learning design) and analyze related performance changes
- minimize the potential of this technology to embed and reproduce existing health inequality and systemic biases
- estimate the generalizability of findings across sites and populations
- enable reproducibility of the findings in the context of a dynamic innovation field and intellectual property protection

Despite indications that some AI-based algorithms now match the accuracy of human experts within preclinical in silico studies[2], there is little high-quality evidence for improved clinician performance or patient outcomes in clinical studies[3,4]. Reasons proposed for this so-called AI chasm[5] are lack of necessary expertise needed for translating a tool into practice, lack of funding available for translation, a general underappreciation of clinical research as a translation mechanism[6] and, more specifically, a disregard for the potential value of the early stages of clinical evaluation and the analysis of human factors[7].

The challenges of early-stage clinical AI evaluation (Box 1) are similar to those of complex interventions, as reported by the Medical Research Council dedicated guidance[1], and surgical innovation, as described by the IDEAL Framework[8,9]. For example, in all three cases, the evaluation needs to consider the potential for iterative modification of the interventions and the characteristics of the operators (or users) performing them. In this regard, the IDEAL framework offers readily implementable and stage-specific recommendations for the evaluation of surgical innovations under development. IDEAL stages 2a and 2b, for example, are described as development and exploratory stages, during which the intervention is refined, operators' learning curves are analyzed and the influence of patient and operator variability on effectiveness are explored prospectively, before large-scale efficacy testing.

Early-stage clinical evaluation of AI systems should also place a strong emphasis on validation of performance and safety, in a similar manner to phase 1 and phase 2 pharmaceutical trials, before efficacy evaluation at scale in phase 3. For example, small changes in the distribution of the underlying data between the algorithm training and clinical evaluation populations (so-called dataset shift) can lead to substantial variation in clinical performance and expose patients to potential unexpected harm[10,11].

Human factors (or ergonomics) evaluations are commonly conducted in safety-critical fields such as aviation, military and energy sectors[12–14]. Their assessments evaluate the effect of a device or procedure on their users' physical and cognitive performance and vice-versa. Human factors, such as usability evaluation, are an integral part of the regulatory process for new medical devices[15,16], and their application to AI-specific challenges is attracting growing attention in the medical literature[17–20]. However, few clinical AI studies have reported on the evaluation of human factors[3], and usability evaluation of related digital health technology is often performed with inconstant methodology and reporting[21].

Other areas of suboptimal reporting of clinical AI studies have also recently been highlighted[3,22], such as implementation environment, user characteristics and selection process, training provided, underlying algorithm identification and disclosure of funding sources. Transparent reporting is necessary for informed study appraisal and to facilitate reproducibility of study results. In a relatively new and dynamic field such as clinical AI, comprehensive reporting is also key to construct a common and comparable knowledge base to build upon.

Guidelines already exist, or are under development, for the reporting of preclinical, in silico studies of AI systems, their offline validation and their evaluation in large comparative studies[23–26]; but there is an important stage of research between these, namely studies focusing on the initial clinical use of AI systems, for which no such guidance currently exists (Fig. 1 and Table 1). This early clinical evaluation provides a crucial scoping evaluation of clinical utility, safety and human factors challenges in live clinical settings. By investigating the potential obstacles to clinical evaluation at scale and informing protocol design, these studies are also important stepping stones toward definitive comparative trials.

To address this gap, we convened an international, multi-stakeholder group of experts in a Delphi exercise to produce the DECIDE-AI reporting guideline. Focusing on AI systems supporting, rather than replacing, human intelligence, DECIDE-AI aims to improve the reporting of studies describing the evaluation of AI-based decision support systems during their early, small-scale implementation in live clinical settings (that is, the supported decisions have an actual effect on patient care). Whereas TRIPOD-AI, STARD-AI, SPIRIT-AI and CONSORT-AI are specific to particular study designs, DECIDE-AI is focused on the evaluation stage and does not prescribe a fixed study design.

Recommendations

Reporting item checklist. The DECIDE-AI guideline should be used for the reporting of studies describing the early-stage live clinical evaluation of AI-based decision support systems, independently of the study design chosen (Fig. 1 and Table 1). Depending on the chosen study design, and if available, authors may also want to complete the reporting according to study-type-specific guidelines (for example, STROBE for cohort studies)[27]. Table 2 presents the DECIDE-AI checklist, comprising the 17 AI-specific reporting items and ten generic reporting items selected by the Consensus Group. Each item comes with an E&E to explain why and how reporting is recommended (Supplementary Appendix 1). A downloadable version of the checklist, designed to help researchers and reviewers check compliance when preparing or reviewing a manuscript, is available as Supplementary Appendix 2. Reporting guidelines are a set of minimum reporting recommendations and not intended to guide research conduct. Although familiarity with DECIDE-AI might be useful to inform some aspects of the design and conduct of studies within the guideline's scope[28], adherence to the guideline alone should not be interpreted as an indication of methodological quality (which is the realm of methodological guidelines and risk of bias assessment tools). With increasingly complex AI interventions and evaluations, it might become challenging to report all the required information within a single primary manuscript, in which case references to the study protocol, open science repositories, related publications and supplementary materials are encouraged.

Discussion

The DECIDE-AI guideline is the result of an international consensus process involving a diverse group of experts spanning a wide range

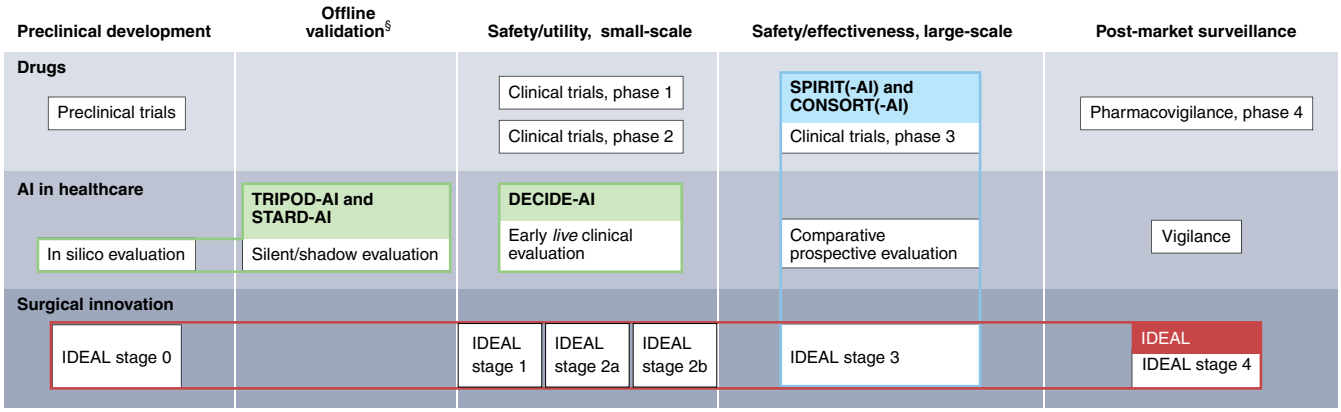

Fig. 1 | Comparison of development pathways for drug therapies, AI in healthcare and surgical innovation. The colored lines represent reporting guidelines, some of which are study design specific (TRIPOD-AI, STARD-AI, SPIRIT/CONSORT and SPIRIT/CONSORT-AI); others are stage specific (DECIDE-AI and IDEAL). Depending on the context, more than one study design can be appropriate for each stage. §Apply only to AI in healthcare.

of professional backgrounds and experience. The level of interest across stakeholder groups and the high response rate among the invited experts speaks to the perceived need for more guidance in the reporting of studies presenting the development and evaluation of clinical AI systems and to the growing value placed on comprehensive clinical evaluation to guide implementation. The emphasis placed on the role of human-in-the-loop decision-making was guided by the Steering Group's belief that AI will, at least in the foreseeable future, augment, rather than replace, human intelligence in clinical settings. In this context, thorough evaluation of the human–computer interaction and the roles played by the human users will be key to realizing the full potential of AI.

The DECIDE-AI guideline is the first stage-specific AI reporting guideline to be developed. This stage-specific approach echoes recognized development pathways for complex interventions[1,8,9,29] and aligns conceptually with proposed frameworks for clinical AI[6,30–32], although no commonly agreed nomenclature or definition has so far been published for the stages of evaluation in this field. Given the current state of clinical AI evaluation, and the apparent deficit in reporting guidance for the early clinical stage, the DECIDE-AI Steering Group considered it important to crystallize current expert opinion into a consensus, to help improve reporting of these studies. Beside this primary objective, the DECIDE-AI guideline will hopefully also support authors during study design, protocol drafting and study registration, by providing them with clear criteria around which to plan their work. As with other reporting guidelines, it is important to note that the overall effect on the standard of reporting will need to be assessed in due course, once the wider community has had a chance to use the checklist and explanatory documents, which is likely to prompt modification and fine-tuning of the DECIDE-AI guideline, based on its real-world use. Although the outcome of this process cannot be pre-judged, there is evidence that the adoption of consensus-based reporting guidelines (such as CONSORT) does, indeed, improve the standard of reporting[33].

The Steering Group paid special attention to the integration of DECIDE-AI within the broader scheme of AI guidelines (for example, TRIPOD-AI, STARD-AI, SPIRIT-AI and CONSORT-AI). It also focused on DECIDE-AI being applicable to all types of decision support modalities (that is, detection, diagnostic, prognostic and therapeutic). The final checklist should be considered as minimum scientific reporting standards and does not preclude reporting additional information, nor are the standards a substitute for other regulatory reporting or approval requirements. The overlap between scientific evaluation and regulatory processes was a core consideration during the development of the DECIDE-AI guideline. Early-stage scientific studies can be used to inform regulatory decisions (for example, based on the stated intended use within the study) and are part of the clinical evidence generation process (for example, clinical investigations). The initial item list was aligned with information commonly required by regulatory agencies, and regulatory considerations are introduced in the E&E paragraphs. However, given the somewhat different focuses of scientific evaluation and regulatory assessment[34], as well as differences between regulatory jurisdictions, it was decided to make no reference to specific regulatory processes in the guideline, nor to define the scope of DECIDE-AI within any particular regulatory framework. The primary focus of DECIDE-AI is scientific evaluation and reporting, for which regulatory documents often provide little guidance.

Several topics led to more intense discussion than others, both during the Delphi process and the Consensus Group discussion. Regardless of whether the corresponding items were included, these represent important issues that the AI and healthcare communities should consider and continue to debate. First, we discussed at length whether users (see glossary of terms) should be considered as study participants. The consensus reached was that users are a key study population, about whom data will be collected (for example, reasons for variation from the AI system recommendation and user satisfaction), and who might logically be consented as study participants and, therefore, should be considered as such. Because user characteristics (for example, experience) can affect intervention efficacy, both patient and user variability should be considered when evaluating AI systems and reported adequately.

Second, the relevance of comparator groups in early-stage clinical evaluation was considered. Most studies retrieved in the literature search described a comparator group (commonly the same group of clinicians without AI support). Such comparators can provide useful information for the design of future large-scale trials (for example, information on the potential effect size). However, comparator groups are often unnecessary at this early stage of clinical evaluation, when the focus is on issues other than comparative efficacy. Small-scale clinical investigations are also usually underpowered to make statistically significant conclusions about efficacy, accounting for both patient and user variability. Moreover, the additional information gained from comparator groups in this context can often be inferred from other sources, such as previous data on unassisted standard of care in the case of the expected effect size. Comparison groups are, therefore, mentioned in item VII but considered optional.

Third, output interpretability is often described as important to increase user and patient trust in the AI system, to contextualize

Table 1 | Overview of existing and upcoming AI reporting guidelines

AI reporting guidelines			
Name	**Stage**	**Study design**	**Comment**
TRIPOD-AI	Preclinical development	**Prediction model evaluation**	Extension of TRIPOD. Used to report prediction models (diagnostic or prognostic) development, validation and updates. Focuses on model performance.
STARD-AI	Preclinical development and offline validation	**Diagnostic accuracy studies**	Extension of STARD. Used to report diagnostic accuracy studies, either at development stage or as an offline validation in clinical settings. Focuses on diagnostic accuracy.
DECIDE-AI	**Early live clinical evaluation**	Various (prospective cohort studies, non-randomized controlled trials, …)[a] with additional features, such as modification of intervention, analysis of pre-specified subgroups or learning curve analysis.	Stand-alone guideline. Used to report the early evaluation of AI systems as an intervention in live clinical settings (small-scale, formative evaluation), independently of the study design and AI system modality (diagnostic, prognostic, therapeutic). Focuses on clinical utility, safety and human factors.
SPIRIT-AI	Comparative prospective evaluation	**Randomized controlled trials (protocol)**	Extension of SPIRIT. Used to report the protocols of randomized controlled trials evaluating AI systems as interventions.
CONSORT-AI	Comparative prospective evaluation	**Randomized controlled trials**	Extension of CONSORT. Used to report randomized controlled trials evaluating AI systems as interventions (large-scale, summative evaluation), independently of the AI system modality (diagnostic, prognostic, therapeutic). Focuses on effectiveness and safety.

Bold font indicates the primary target of the guidelines, either a specific stage or a specific study design. [a]Although existing reporting guidelines exist for some of these study designs (for example, STROBE for cohort studies), none covers all the core aspects of AI system early-stage evaluation, and none would fit all possible study designs; DECIDE-AI was, therefore, developed as a new stand-alone reporting guideline for these early, live, clinical AI studies.

the system's outputs within the broader clinical information environment[19] and potentially for regulatory purposes[35]. However, some experts argued that an output's clinical value may be independent of its interpretability and that the practical relevance of evaluating interpretability is still debatable[36,37]. Furthermore, there is currently no generally accepted way of quantifying or evaluating interpretability. For this reason, the Consensus Group decided not to include an item on interpretability at the current time.

Fourth, the notion of users' trust in the AI system and its evolution with time were discussed. As users accumulate experience with, and receive feedback from, the real-world use of AI systems, they will adapt their level of trust in its recommendations. Whether appropriate or not, this level of trust will influence, as recently demonstrated by McIntosh et al.[38], how much effect the systems have on the final decision-making and, therefore, influence the overall clinical performance of the AI system. Understanding how trust evolves is essential for planning user training and determining the optimal timepoints at which to start data collection in comparative trials. However, as for interpretability, there is currently no commonly accepted way to measure trust in the context of clinical AI. For this reason, the item about user trust in the AI system was not included in the final guideline. The fact that interpretability and trust were not included highlights the tendency of consensus-based guidelines development toward conservatism, because only widely agreed-upon concepts reach the level of consensus needed for inclusion. However, changes of focus in the field, as well as new methodological development, can be integrated into subsequent guideline iterations. From this perspective, the issues of interpretability and trust are far from irrelevant to future AI evaluations, and their exclusion from the current guideline reflects less a lack of interest than a need for further research into how we can best operationalize these metrics for the purposes of evaluation in AI systems.

Fifth, the notion of modifying the AI system (the intervention) during the evaluation received mixed opinions. During comparative trials, changes made to the intervention during data collection are questionable unless the changes are part of the study protocol; some authors even consider them as impermissible, on the basis that they would make valid interpretation of study results difficult or impossible. However, the objectives of early clinical evaluation are often not to make definitive conclusions on effectiveness. Iterative design–evaluation cycles, if performed safely and reported transparently, offer opportunities to tailor an intervention to its users and beneficiaries and augment chances of adoption of an optimized, fixed version during later summative evaluation[8,9,39,40].

Sixth, several experts noted the benefit of conducting human factors evaluation before clinical implementation and considered that, therefore, human factors should be reported separately. However, even robust preclinical human factors evaluation will not reliably characterize all the potential human factors issues that might arise during the use of an AI system in a live clinical environment, warranting a continued human factors evaluation at the early stage of clinical implementation. The Consensus Group agreed that human factors play a fundamental role in AI system adoption in clinical settings at scale and that the full appraisal of an AI system's clinical utility can happen only in the context of its clinical human factors evaluation.

Finally, several experts raised concerns that the DECIDE-AI guideline prescribes an evaluation that is too exhaustive to be reported within a single manuscript. The Consensus Group acknowledged the breadth of topics covered and the practical implications. However, reporting guidelines aim to promote transparent reporting of studies rather than mandating that every aspect covered by an item must have been evaluated within the studies. For example, if a learning curves evaluation has not been performed, then fulfilment of item 14b would be to simply state that this was not done, with an accompanying rationale. The Consensus Group agreed that appropriate AI evaluation is a complex endeavour necessitating the interpretation of a wide range of data, which should be presented together as far as possible. It was also felt that thorough evaluation of AI systems should not be limited by a word count and that publications reporting on such systems might benefit from special formatting requirements in the future. The information required by

Table 2 | DECIDE-AI checklist

Item number	Theme	Recommendation
1–17	AI-specific reporting items	
1–X	Generic reporting items	
Title and abstract		
1	Title	Identify the study as early clinical evaluation of a decision support system based on AI or machine learning, specifying the problem addressed.
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 and secondary outcomes, key safety endpoints, human factors evaluated, main results and conclusions.
Introduction		
2	Intended use	a) Describe the targeted medical condition(s) and problem(s), including the current standard practice, and the intended patient population(s).
		b) Describe the intended users of the AI system, its planned integration in the care pathway, and the potential effect, including patient outcomes, that it is intended to have.
II	Objectives	State the study objectives.
Methods		
III	Research governance	Provide a reference to any study protocol, study registration number, and ethics approval.
3	Participants	a) 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.
		b) Describe how users were recruited, stating the inclusion and exclusion criteria, and how the intended number of recruited users was decided.
		c) Describe steps taken to familiarize the users with the AI system, including any training received before the study.
4	AI system	a) Briefly describe the AI system, specifying its version and 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.
		b) 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.
		c) Describe the AI system outputs and how they were presented to the users (an image may be useful).
5	Implementation	a) Describe the settings in which the AI system was evaluated.
		b) Describe the clinical workflow/care pathway in which the AI system was evaluated, the timing of its use, and how the final supported decision was reached and by whom.
IV	Outcomes	Specify the primary and secondary outcomes measured.
6	Safety and errors	a) Provide a description of how significant errors/malfunctions were defined and identified.
		b) Describe how any risks to patient safety or instances of harm were identified, analyzed, and minimized.
7	Human factors	Describe the human factors tools, methods or frameworks used, the use cases considered, and the users involved.
V	Analysis	Describe the statistical methods by which the primary and secondary outcomes were analyzed, as well as any pre-specified additional analyses, including subgroup analyses and their rationale.
8	Ethics	Describe whether specific methodologies were used to fulfil an ethics-related goal (such as algorithmic fairness) and their rationale.
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.
Results		
9	Participants	a) Describe the baseline characteristics of the patients included in the study and report on input data missingness.
		b) Describe the baseline characteristics of the users included in the study.
10	Implementation	a) 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.
		b) Report any significant changes to the clinical workflow or care pathway caused by the AI system.
VII	Main results	Report on the pre-specified outcomes, including outcomes for any comparison group if applicable.
VIII	Subgroups analysis	Report on the differences in the main outcomes according to the pre-specified subgroups.
11	Modifications	Report any changes made to the AI system or its hardware platform during the study. Report the timing of these modifications, the rationale for each, and any changes in outcomes observed after each of them.

Continued

Table 2 | DECIDE-AI checklist (continued)

Item number	Theme	Recommendation
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's recommendations.
13	Safety and errors	a) 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 effects on patient care.
		b) Report on any risks to patient safety or observed instances of harm (including indirect harm) identified during the study.
14	Human factors	a) Report on the usability evaluation, according to recognized standards or frameworks.
		b) Report on the user learning curves evaluation.
Discussion		
15	Support for intended use	Discuss whether the results obtained support the intended use of the AI system in clinical settings.
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.
IX	Strengths and limitations	Discuss the strengths and limitations of the study.
Statements		
17	Data availability	Disclose if and how data and relevant code are available.
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.

AI-specific items are numbered in Arabic numerals; generic items are numbered in Roman numerals.

several items might already be reported in previous studies or in the study protocol, which could be cited rather than described in full again. The use of references, online supplementary materials and open-access repositories (for example, Open Science Framework (OSF)) is recommended to allow the sharing and connecting of all required information within one main published evaluation report.

Our work has several limitations that should be considered. First, the issue of potential biases, which apply to any consensus process, must be considered. These include anchoring or participant selection biases[41]. The research team tried to mitigate bias through the survey design, using open-ended questions analyzed through a thematic analysis, and by adapting the expert recruitment process, but it is unlikely that it was eliminated entirely. Despite an aim for geographical diversity and several actions taken to foster it, representation was skewed toward Europe and, more specifically, the United Kingdom. This could be explained, in part, by the following factors: a likely selection bias in the Steering Group's expert recommendations; a higher interest in our open invitation to contribute coming from European/United Kingdom scientists (25 of 30 experts approaching us, 83%); and a lack of control over the response rate and self-reported geographical location of participating experts. Considerable attention was also paid to diversity and balance among stakeholder groups, even though clinicians and engineers were the most represented, partly due to the profile of researchers who contacted us spontaneously after the public announcement of the project. Stakeholder group analyses were performed to identify any marked disagreements from underrepresented groups. Finally, as also noted by the authors of the SPIRIT-AI and CONSORT-AI guidelines[25,26], few examples of studies reporting on the early-stage clinical evaluation of AI tools were available at the time that we started developing the DECIDE-AI guideline. This might have affeced the exhaustiveness of the initial item list created from literature review. However, the wide range of stakeholders involved and the design of the first round of Delphi allowed identification of several additional candidate items, which were added in the second iteration of the item list.

The introduction of AI into healthcare needs to be supported by sound, robust and comprehensive evidence generation and reporting. This is essential both to ensure the safety and efficacy of AI systems and to gain the trust of patients, practitioners and purchasers, so that this technology can realize its full potential to improve patient care. The DECIDE-AI guideline aims to improve the reporting of early-stage live clinical evaluation of AI systems, which lays the foundations for both larger clinical studies and later widespread adoption.

Methods

The DECIDE-AI guideline was developed through an international expert consensus process and in accordance with the EQUATOR Network's recommendations for guideline development[42]. A Steering Group was convened to oversee the guideline development process. Its members were selected to cover a broad range of expertise and ensure a seamless integration with other existing guidelines. We conducted a modified Delphi process[43], with two rounds of feedback from participating experts and one virtual consensus meeting. The project was reviewed by the University of Oxford Central University Research Ethics Committee (approval R73712/RE003) and registered with the EQUATOR Network. Informed consent was obtained from all participants in the Delphi process and consensus meeting.

Initial item list generation. An initial list of candidate items was developed based on expert opinion informed by (1) a systematic literature review focusing on the evaluation of AI-based diagnostic decision support systems[3]; (2) an additional literature search about existing guidance for AI evaluation in clinical settings (search strategy available on the OSF[44]); (3) literature recommended by Steering Group members[19,22,45–49]; and (4) institutional documents[50–53].

Expert recruitment. Experts were recruited through five different channels: (1) invitation to experts recommended by the Steering Group; (2) invitation to authors of the publications identified through

Box 2 | Glossary of terms

AI system	Decision support system incorporating AI and consisting of (1) the AI or machine learning algorithm; (2) the supporting software platform; and (3) the supporting hardware platform
AI system version	Unique reference for the form of the AI system and the state of its components at a single point in time. Allows for tracking changes to the AI system over time and comparing between different versions.
Algorithm	Mathematical model responsible for learning from data and producing an output.
AI	Science of developing computer systems which can perform tasks normally requiring human intelligence'[26]
Bias	Systematic difference in treatment of certain objects, people, or groups in comparison to others'[58]
Care pathway	Series of interactions, investigations, decision-making and treatments experienced by patients in the course of their contact with a healthcare system for a defined reason
Clinical	Relating to the observation and treatment of actual patients rather than in silico or scenario-based simulations
Clinical evaluation	Set of ongoing activities, analyzing clinical data and using scientific methods, to evaluate the clinical performance, effectiveness and/or safety of an AI system, when used as intended[50]
Clinical investigation	Study performed on one or more human subjects to evaluate the clinical performance, effectiveness and/or safety of an AI system[59]. This can be performed in any setting (for example, community, primary care and hospital).
Clinical workflow	Series of tasks performed by healthcare professionals in the exercise of their clinical duties
Decision support system	System designed to support human decision-making by providing person-specific and situation-specific information or recommendations to improve care or enhance health
Exposure	State of being in contact with, and having used, an AI system or similar digital technology
Human–computer interaction	Bi-directional influence between human users and digital systems through a physical and conceptual interface.
Human factors	Also called ergonomics. 'The scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimise human well-being and overall system performance' (International Ergonomics Association).
Indication for use	Situation and reason (medical condition, problem and patient group) where the AI system should be used
In silico evaluation	Evaluation performed via computer simulation outside the clinical settings
Intended use	Use for which an AI system is intended, as stated by its developers, and which serves as the basis for its regulatory classification. The intended use includes aspects of the targeted medical condition, patient population, user population, use environment and mode of action.
Learning curves	Graphical plotting of user performance against experience[60]. By extension, analysis of the evolution of user performance with a task as exposure to the task increases. The measure of performance often uses other context-specific metrics as a proxy.
Live evaluation	Evaluation under actual clinical conditions, in which the decisions made have a direct effect on patient care. As opposed to 'offline' or 'shadow mode' evaluation where the decisions do not have a direct effect on patient care.
Machine learning	'Field of computer science concerned with the development of models/algorithms that can solve specific tasks by learning patterns from data, rather than by following explicit rules. It is seen as an approach within the field of AI'[26].
Participant	Subject of a research study on whom data will be collected and from whom consent is obtained (or waived). The DECIDE-AI guideline considers that both patients and users can be participants.
Patient	Person (or the digital representation of this person) receiving healthcare attention or using health services and who is the subject of the decision made with the support of the AI system. Note: DECIDE-AI uses the term 'patient' pragmatically to simplify the reading of the guideline. Strictly speaking, a person with no health conditions who is the subject of a decision made about them by an AI-based decision support tool to improve their health and well-being or for a preventative purpose is not necessarily a 'patient' per se.
Patient involvement in research	Research carried out 'with' or 'by' patients or members of the public rather than 'to', 'about' or 'for' them (adapted from the INVOLVE definition of 'Public Involvement').
Standard practice	Usual care currently received by the intended patient population for the targeted medical condition and problem. This may not necessarily be synonymous with the state-of-the-art practice.
Usability	'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'[61].
User	Person interacting with the AI system to inform their decision-making. This person could be a healthcare professional or a patient.

The definitions provided pertain to the specific context of DECIDE-AI and the use of the terms in the guideline. They are not necessarily generally accepted definitions and might not always be fully applicable to other areas of research.

the initial literature searches; (3) call to contribute published in a commentary article in a medical journal[7]; (4) consideration of any expert contacting the Steering Group on their own initiative; and (5) invitation to experts recommended by the Delphi participants (snowballing). Before starting the recruitment process, 20 target stakeholder groups were defined, namely: administrators/hospital management, allied health professionals, clinicians, engineers/computer scientists, entrepreneurs, epidemiologists, ethicists, funders, human

factors specialists, implementation scientists, journal editors, methodologists, patient representatives, payers/commissioners, policymakers/official institution representatives, private sector representatives, psychologists, regulators, statisticians and trialists.

One hundred thirty-eight experts agreed to participate in the first round of Delphi, of whom 123 (89%) completed the questionnaire (83 identified from Steering Group recommendations, 12 from their publications, 21 from contacting the Steering Group on own initiative and seven through snowballing). One hundred sixty-two experts were invited to take part in the second round of Delphi, of whom 138 completed the questionnaire (85%). One hundred ten had also completed the first round (continuity rate of 89%)[54], and 28 were new participants. The participating experts represented 18 countries and spanned all 20 of the defined stakeholder groups (Supplementary Note 1 and Supplementary Tables 1 and 2).

Delphi process. The Delphi surveys were designed and distributed via the REDCap web application[55,56]. The first round consisted of four open-ended questions on aspects viewed by the Delphi participants as necessary to be reported during early-stage clinical evaluation. The participating experts were then asked to rate, on a 1–9 scale, the importance of items in the initial list proposed by the research team. Ratings of 1–3 on the scale were defined as 'not important', 4–6 as 'important but not critical' and 7–9 as 'important and critical'. Participants were also invited to comment on existing items and to suggest new items. An inductive thematic analysis of the narrative answers was performed independently by two reviewers (B.V. and M.N.), and conflict was resolved by consensus[57]. The themes identified were used to correct any omissions in the initial list and to complement the background information about proposed items. Summary statistics of the item scores were produced for each stakeholder group by calculating the median score, the interquartile range (IQR) and the percentage of participants scoring an item 7 or higher, as well as 3 or lower, which were the pre-specified inclusion and exclusion cutoffs, respectively. A revised item list was developed based on the results of the first round.

In the second round, the participants were shown the results of the first round and invited to rate and comment on the items in the revised list. The detailed survey questions of the two rounds of Delphi can be found on the OSF[44]. All analyses of item scores and comments were performed independently by two members of the research team (B.V. and M.N.) using NVivo (QSR International Pty Ltd., version 1.0) and Python (Python Software Foundation, version 3.8.5). Conflicts were resolved by consensus.

The initial item list contained 54 items. One hundred twenty sets of responses were included in the analysis of the first round of Delphi (one set of responses was excluded due to a reasonable suspicion of scale inversion, two due to completion after the deadline). The first round yielded 43,986 words of free text answers to the four initial open-ended questions, 6,419 item scores, 228 comments and 64 proposals for new items. The thematic analysis identified 109 themes. In the revised list, nine items remained unchanged, 22 were reworded/completed, 21 were reorganized (merged/split, becoming 13 items), two items were dropped and nine new items were added, for a total of 53 items. The two items dropped were related to health economic assessment. They were the only two items with a median score below 7 (median: 6, IQR: 2–9 for both) and received many comments describing them as an entirely separate aspect of evaluation. The revised list was reorganized into items and subitems. One hundred thirty-six sets of answers were included in the analysis of the second round of Delphi (one set of answers was excluded due to lack of consideration for the questions, one due to completion after the deadline). The second round yielded 7,101 item scores and 923 comments. The results of the thematic analysis and the initial and revised item lists, as well as per-item narrative and graphical summaries of the feedback received in both rounds, can be found on the OSF[44].

Consensus meeting. A virtual consensus meeting was held on three occasions between 14 and 16 June 2021 to debate and agree to the content and wording of the DECIDE-AI reporting guideline. The 16 members of the Consensus Group (Supplementary Note 1 and Supplementary Table 2a,b) were selected to ensure a balanced representation of the key stakeholder groups as well as geographic diversity. All items from the second round of Delphi were discussed and voted on during the consensus meeting. For each item, the results of the Delphi process were presented to the Consensus Group members, and a vote was carried out anonymously using the Vevox online application (https://www.vevox.com). A pre-specified cutoff of 80% of the Consensus Group members (excluding blank votes and abstentions) was necessary for an item to be included. To highlight the new, AI-specific reporting items, the Consensus Group divided the guidelines into two item lists: an AI-specific items list, which represents the main novelty of the DECIDE-AI guideline, and a second list of generic reporting items, which achieved high consensus but are not AI specific and could apply to most types of studies. The Consensus Group selected 17 items (made of 28 subitems in total) for inclusion in the AI-specific list and ten items for inclusion in the generic reporting item list. A summary of the Consensus Group votes can be found in Supplementary Table 3.

Qualitative evaluation. The drafts of the guideline and of the E&E sections were sent for qualitative evaluation to a group of 16 selected experts with experience in AI system implementation or in the peer-reviewing of literature related to AI system evaluation (Supplementary Note 1), all of whom were independent of the Consensus Group. These 16 experts were asked to comment on the clarity and applicability of each AI-specific item, using a custom form (available on the OSF[44]). Item wording amendments and modifications to the E&E sections were conducted based on the feedback from the qualitative evaluation, which was independently analyzed by two reviewers (B.V. and M.N.) and with conflicts resolved by consensus. A glossary of terms (Box 2) was produced to clarify key concepts used in the guideline. The Consensus Group approved the final item lists, including any changes made during the qualitative evaluation. Supplementary Figs. 1 and 2 provide graphical representations of the two item lists' (AI-specific and generic) evolution.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data generated during this study (pseudonymized where necessary) are available upon justified request to the research team and for a duration of 3 years after publication of this manuscript. Translation of these guidelines into different languages is welcomed and encouraged, as long as the authors of the original publication are included in the process and resulting publication.

Code availability

All codes produced for data analysis during this study are available upon justified request to the research team and for a duration of 3 years after publication of this manuscript.

Received: 20 November 2021; Accepted: 3 March 2022;
Published online: 18 May 2022

References

1. Skivington, K. et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *Br. Med. J.* **374**, n2061 (2021).

2. Liu, X. et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *Lancet Digit. Health* **1**, e271–e297 (2019).
3. Vasey, B. et al. Association of clinician diagnostic performance with machine learning-based decision support systems: a systematic review. *JAMA Netw. Open* **4**, e211276 (2021).
4. Freeman, K. et al. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *Br. Med. J.* **374**, n1872 (2021).
5. Keane, P. A. & Topol, E. J. With an eye to AI and autonomous diagnosis. *NPJ Digital Med.* **1**, 40 (2018).
6. McCradden, M. D., Stephenson, E. A. & Anderson, J. A. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat. Med.* **26**, 1325–1326 (2020).
7. Vasey, B. et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat. Med.* **27**, 186–187 (2021).
8. McCulloch, P. et al. No surgical innovation without evaluation: the IDEAL recommendations. *Lancet* **374**, 1105–1112 (2009).
9. 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).
10. Finlayson, S. G. et al. The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* **385**, 283–286 (2021).
11. Subbaswamy, A. & Saria, S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* **21**, 345–352 (2020).
12. Kapur, N., Parand, A., Soukup, T., Reader, T. & Sevdalis, N. Aviation and healthcare: a comparative review with implications for patient safety. *JRSM Open* **7**, 2054270415616548 (2015).
13. Corbridge, C., Anthony, M., McNeish, D. & Shaw, G. A new UK defence standard for human factors integration (HFI). *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **60**, 1736–1740 (2016).
14. Stanton, N. A., Salmon, P., Jenkins, D. & Walker, G. *Human Factors in the Design and Evaluation of Central Control Room Operations* (CRC Press, 2009).
15. US Food and Drug Administration (FDA). Applying human factors and usability engineering to medical device: guidance for industry and Food and Drug Administration staff. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/applying-human-factors-and-usability-engineering-medical-devices (2016).
16. Medicines & Healthcare products Regulatory Agency (MHRA). Guidance on applying human factors and usability engineering to medical devices including drug-device combination products in Great Britain. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/970563/Human-Factors_Medical-Devices_v2.0.pdf (2021).
17. Asan, O. & Choudhury, A. Research trends in artificial intelligence applications in human factors health care: mapping review. *JMIR Hum. Factors* **8**, e28236 (2021).
18. Felmingham, C. M. et al. The importance of incorporating human factors in the design and implementation of artificial intelligence for skin cancer diagnosis in the real world. *Am. J. Clin. Dermatol.* **22**, 233–242 (2021).
19. Sujan, M. et al. Human factors challenges for the safe use of artificial intelligence in patient care. *BMJ Health Care Inform.* **26**, e100081 (2019).
20. Sujan, M., Baber, C., Salmon, P., Pool, R. & Chozos, N. Human factors and ergonomics in healthcare AI. https://www.researchgate.net/publication/354728442_Human_Factors_and_Ergonomics_in_Healthcare_AI (2021).
21. Wronikowska, M. W. et al. Systematic review of applied usability metrics within usability evaluation methods for hospital electronic healthcare record systems. *J. Eval. Clin. Pract.* **27**, 1403–1416 (2021).
22. Nagendran, M. et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *Br. Med. J.* **368**, m689 (2020).
23. Collins, G. S. & Moons, K. G. M. Reporting of artificial intelligence prediction models. *Lancet* **393**, 1577–1579 (2019).
24. Sounderajah, V. et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. *Nat. Med.* **26**, 807–808 (2020).
25. Cruz Rivera, S. et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat. Med.* **26**, 1351–1363 (2020).
26. 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).
27. von Elm, E. et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Br. Med. J.* **335**, 806–808 (2007).
28. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **372**, n71 (2021).
29. Sedrakyan, A. et al. IDEAL-D: a rational framework for evaluating and regulating the use of medical devices. *Br. Med. J.* **353**, i2372 (2016).
30. Park, Y. et al. Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* **3**, 326–331 (2020).
31. Higgins, D. & Madai, V. I. From bit to bedside: a practical framework for artificial intelligence product development in healthcare. *Adv. Intell. Syst.* **2**, 2000052 (2020).
32. Sendak, M. P. et al. A path for translation of machine learning products into healthcare delivery. *Eur. Med. J.* https://www.emjreviews.com/innovations/article/a-path-for-translation-of-machine-learning-products-into-healthcare-delivery/ (2020).
33. Moher, D., Jones, A., Lepage, L. & CONSORT Group. Use of the CONSORT statement and quality of reports of randomized trials: a comparative before-and-after evaluation. *J. Am. Med. Assoc.* **285**, 1992–1995 (2001).
34. Park, S. H. Regulatory approval versus clinical validation of artificial intelligence diagnostic tools. *Radiology* **288**, 910–911 (2018).
35. 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).
36. Lipton, Z. C. The mythos of model interpretability. *Commun. ACM* **61**, 36–43 (2018).
37. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Health* **3**, e745–e750 (2021).
38. McIntosh, C. et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nat. Med.* **27**, 999–1005 (2021).
39. International Organization for Standardization. Ergonomics of human–system interaction—part 210: human-centred design for interactive systems. https://www.iso.org/standard/77520.html (2019).
40. Norman, D. A. *User Centered System Design* (CRC Press, 1986).
41. Winkler, J. & Moser, R. Biases in future-oriented Delphi studies: a cognitive perspective. *Technol. Forecast. Soc. Change* **105**, 63–76 (2016).
42. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for developers of health research reporting guidelines. *PLoS Med.* **7**, e1000217 (2010).
43. Dalkey, N. & Helmer, O. An experimental application of the DELPHI method to the use of experts. *Manage. Sci.* **9**, 458–467 (1963).
44. Vasey, B., Nagendran, M. & McCulloch, P. DECIDE-AI 2022. https://doi.org/10.17605/OSF.IO/TP9QV (2022).
45. Vollmer, S. et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *Br. Med. J.* **368**, l6927 (2020).
46. Bilbro, N. A. et al. The IDEAL reporting guidelines: a Delphi consensus statement stage specific recommendations for reporting the evaluation of surgical innovation. *Ann. Surg.* **273**, 82–85 (2021).
47. Morley, J., Floridi, L., Kinsey, L. & Elhalal, A. From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Sci. Eng. Ethics* **26**, 2141–2168 (2019).
48. Xie, Y. et al. Health economic and safety considerations for artificial intelligence applications in diabetic retinopathy screening. *Transl. Vis. Sci. Technol.* **9**, 22 (2020).
49. Norgeot, B. et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* **26**, 1320–1324 (2020).
50. IMDRF Medical Device Clinical Evaluation Working Group. Clinical Evaluation. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n56.pdf (2019).
51. IMDRF Software as Medical Device (SaMD) Working Group. 'Software as a medical device': possible framework for risk categorization and corresponding considerations. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf (2014).
52. National Institute for Health and Care Excellence (NICE). Evidence standards framework for digital health technologies. https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies (2019).
53. High-Level Independent Group on Artificial Intelligence (AI HLEG). Ethics guidelines for trustworthy AI. European Commission. Vol. 32. https://ec.europa.eu/digital (2019).
54. Boel, A., Navarro-Compán, V., Landewé, R. & van der Heijde, D. Two different invitation approaches for consecutive rounds of a Delphi survey led to comparable final outcome. *J. Clin. Epidemiol.* **129**, 31–39 (2021).
55. Harris, P. A. et al. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J. Biomed. Inform.* **42**, 377–381 (2009).
56. Harris, P. A. et al. The REDCap consortium: building an international community of software platform partners. *J. Biomed. Inform.* **95**, 103208 (2019).
57. Nowell, L. S., Norris, J. M., White, D. E. & Moules, N. J. Thematic analysis: striving to meet the trustworthiness criteria. *Int. J. Qual. Methods* **16**, 1609406917733847 (2017).

58. International Organization for Standardization. Information technology—artificial intelligence (AI)—bias in AI systems and AI aided decision making. https://www.iso.org/standard/77607.html (2021).
59. IMDRF Medical Device Clinical Evaluation Working Group. Clinical Investigation. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n57.pdf (2019).
60. Hopper, A. N., Jamison, M. H. & Lewis, W. G. Learning curves in surgical practice. *Postgrad. Med. J.* **83**, 777–779 (2007).
61. International Organization for Standardization. Ergonomics of human–system interaction—part 11: usability: definitions and concepts. https://www.iso.org/standard/63500.html (2018).

Acknowledgements

The authors would like to thank all Delphi participants and experts who participated in the guideline qualitative evaluation. B.V. would also like to thank B. Beddoe (Sheffield Teaching Hospital), N. Bilbro (Maimonides Medical Center), N. Marlow (Oxford University Hospitals), E. Taylor (Nuffield Department of Surgical Sciences, University of Oxford) and S. Ursprung (Department for Radiology, Tübingen University Hospital) for their support in the initial stage of the project. This work was supported by the IDEAL Collaboration. B.V. is funded by a Berrow Foundation Lord Florey scholarship. M.N. is supported by the UKRI CDT in AI for Healthcare (http://ai4health.io; grant P/S023283/1). D.C. receives funding from the Wellcome Trust, AstraZeneca, RCUK and GlaxoSmithKline. G.S.C. is supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK (program grant C49297/A27294). M.I. is supported by a Maimonides Medical Center Research fellowship. X.L. receives funding from the Wellcome Trust, the National Institute of Health Research/NHSX/Health Foundation, the Alan Turing Institute, the MHRA and NICE. B.A.M. is a fellow of the Alan Turing Institute, supported by EPSRC grant EP/N510129/, and holds a Wellcome Trust-funded honorary post at University College London for the purposes of carrying out independent research. M.M. receives funding from the Dalla Lana School of Public Health and the Leong Centre for Healthy Children. J.O. is employed by the Medicines and Healthcare products Regulatory Agency, which is the competent authority responsible for regulating medical devices and medicines in the United Kingdom. Elements of the work relating to the regulation of AI as a medical device are funded by grants from NHSX and the Regulators' Pioneer Fund (Department for Business, Energy and Industrial Strategy). S.S. receives grants from the National Science Foundation, the American Heart Association, the National Institutes of Health and the Sloan Foundation. D.S.W.T. is supported by the National Medical Research Council, Singapore (NMRC/HSRG/0087/2018;MOH-000655-00), the National Health Innovation Centre, Singapore (NHIC-COV19-2005017), the SingHealth Fund Limited Foundation (SHF/HSR113/2017), the Duke-NUS Medical School (Duke-NUS/RSF/2021/0018;05/FY2020/EX/15-A58) and the Agency for Science, Technology and Research (A20H4g2141; H20C6a0032). P. Watkinson is supported by the NIHR Biomedical Research Centre, Oxford, and holds grants from the NIHR and Wellcome. P. McCulloch receives grants from Medtronic (unrestricted educational grant to Oxford University for the IDEAL Collaboration) and the Oxford Biomedical Research Centre. The views expressed in this guideline are those of the authors, Delphi participants and experts who participated in the qualitative evaluation of the guideline. These views do not necessarily reflect those of their institutions or funders.

Author contributions

B.V., M.N. and P. McCulloch designed the study. B.V. and M.I. conducted the literature searches. Members of the DECIDE-AI Steering Group (B.V., D.C., G.S.C., A.K.D., L.F., B.G., X.L., P. Mathur, L.M., S.S., P. Watkinson and P. McCulloch) provided methodological input and oversaw the conduct of the study. B.V. and M.N. conducted the thematic analysis and Delphi rounds analysis and produced the Delphi round summaries. Members of the DECIDE-AI Consensus Group (B.V., G.S.C., S.P., B.G., X.L., B.A.M., P. Mathur., M.M., L.M., J.O., C.R., S.S., D.S.W.T., W.W., P. Wheatstone and P. McCulloch) selected the final content and wording of the guidelines. B.C. chaired the consensus meeting. B.V., M.N. and B.C. drafted the final manuscript and E&E sections. All authors reviewed and commented on the final manuscript and E&E sections. All members of the DECIDE-AI expert group collaborated in the development of the DECIDE-AI guidelines by participating in the Delphi process, the qualitative evaluation of the guidelines or both.

Competing interests

M.N. consults for Cera Care, a technology-enabled homecare provider. B.C. was a Non-Executive Director of the UK Medicines and Healthcare products Regulatory Agency (MHRA) from September 2015 until 31 August 2021. D.C. receives consulting fees from Oxford University Innovation, Biobeats and Sensyne Health and has an advisory role with Bristol Myers Squibb. B.G. has received consultancy and research grants from Philips NV and Edwards Lifesciences LLC and is owner and board member of Healthplus.ai BV and its subsidiaries. X.L. has advisory roles with the National Screening Committee UK, the WHO/ITU focus group for AI in health and the AI in Health and Care Award Evaluation Advisory Group (NHSX, AAC). P. Mathur is the co-founder of BrainX LLC and BrainX Community LLC. M.M. reports consulting fees from AMS Healthcare and honoraria from the Osgoode Law School and the Toronto Pain Institute. L.M. is director and owner of Morgan Human Systems. J.O. holds an honorary post as an Associate of Hughes Hall, University of Cambridge. C.R. is an employee of HeartFlow Inc., including salary and equity. S.S. has received honoraria from several universities and pharmaceutical companies for talks on digital health and AI. S.S. has advisory roles in Child Health Imprints, Duality Tech, Halcyon Health and Bayesian Health. S.S. is on the board of Bayesian Health. This arrangement has been reviewed and approved by Johns Hopkins in accordance with its conflict of interest policies. D.S.W.T. holds patents linked to AI-driven technologies and is a co-founder and equity holder of EyRIS Pte Ltd. P. Watkinson declares grants, consulting fees and stocks from Sensyne Health and holds patents linked to AI-driven technologies. P. McCulloch has an advisory role for WEISS International and the technology incubator PhD program at University College London. B.V., G.S.C., A.K.D., L.F., M.I., B.A.M., S.D., P. Wheatstone and W.W. have no further conflicts of interest to declare.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-01772-9.

Correspondence and requests for materials should be addressed to Baptiste Vasey.

Peer review information *Nature Medicine* thanks Alejandro Berlin, Rahul Deo, Isabelle Boutron and Leo Anthony Celi for their contribution to the peer review of this work. Javier Carmona was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

nature portfolio

Corresponding author(s): Baptiste Vasey

Last updated by author(s): Feb 9, 2022

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	See line 205 and lines 238-239
Data analysis	See lines 240-242

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data generated during this study (pseudonymised where necessary) are available upon justified request to the research team and for a duration of three years after publication of this manuscript. Translation of this guideline into different languages is welcomed and encouraged, as long as the authors of the original publication are included in the process and resulting publication.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	See lines 196-201.
Data exclusions	See lines 244-245 and lines 270-271.
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Supplementary tables 1a-b and 2a-b
Recruitment	See lines 175-195
Ethics oversight	See lines 164-167

Note that full information on the approval of the study protocol must also be provided in the manuscript.