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

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

[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, UK
[14] Moorfields Eye Hospital NHS Foundation Trust, London, UK
[15] Healthplus.ai-R&D B.V., 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, OH, USA
[20] The Hospital for Sick Children, Toronto, Canada
[21] Dalla Lana School of Public Health, University of Toronto, Canada
[22] Morgan Human Systems Ltd, Shrewsbury, UK
[23] The 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, SG
[28] Duke-NUS Medical School, National University of Singapore, Singapore, SG
[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] A full list of members and their affiliations appears in the Supplementary Information.
* Corresponding author: Baptiste Vasey, Nuffield Department of Surgical Sciences, University of Oxford, Headington, Oxford OX3 9DU, United Kingdom (baptiste.vasey@nds.ox.ac.uk).

ABSTRACT

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 multistakeholder, 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 analyse expert opinion on the reporting of early clinical evaluation of AI systems. Experts were recruited from 20 predefined 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. 123 experts participated in the first round of Delphi, 138 in the second, 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 10 generic reporting items, with an E&E paragraph provided for each. Through consultation and consensus with a range of stakeholders, we have 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.

Main

The prospect of improved clinical outcomes and more efficient health systems has fuelled 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 between 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]. Despite indications that some AI-based algorithms now match the accuracy of human experts within pre-clinical *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 (see 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/2b, for example, are described as development and exploratory stages, during which the intervention is refined, operators' learning curves analysed, and the influence of patient and operator variability on effectiveness are explored prospectively, prior to 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 I and II pharmaceutical trials, before efficacy evaluation at scale in phase III. 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 significant 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, the military and energy sectors[12–14]. Their assessments evaluate the impact 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 report 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 for their evaluation in large comparative studies[23–26]; but there is an important stage of research between these, namely studies focussing on the initial clinical use of AI systems, for which no such guidance currently exists (see Figure 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, multistakeholder 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 (i.e. the supported decisions have an actual impact 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.

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[27]. 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[28], 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 number 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: (i) a systematic literature review focusing on the evaluation of AI-based diagnostic decision support systems[3], (ii) an additional literature search about existing guidance for AI evaluation in clinical settings (search strategy available on the Open Science Framework[29]), (iii) literature recommended by Steering Group members[19,22,30–34], and (iv) institutional documents[35–39].

Expert recruitment

Experts were recruited through five different channels: (i) invitation to experts recommended by the Steering Group, (ii) invitation to authors of the publications identified through the initial literature searches, (iii) call to contribute published in a commentary article in a

medical journal[7], (iv) consideration of any expert contacting the Steering Group of their own initiative, and (v) 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, policy makers/official institution representatives, private sector representatives, psychologists, regulators, statisticians, and trialists.

138 experts agreed to participate in the first round of Delphi, of whom 123 (89%) completed the questionnaire (83 identified from Steering Group recommendation, 12 from their publications, 21 contacting the Steering Group from of own initiative, and seven through snowballing). 162 experts were invited to take part in the second round of Delphi, of whom 138 completed the questionnaire (85%). 110 had also completed the first round (continuity rate of 89%)[40] and 28 were new participants. The participating experts represented 18 countries and spanned all 20 of the defined stakeholder groups (see Suppl. notes 1 and Suppl. tables 1 and 2).

Delphi process

The Delphi surveys were designed and distributed via the REDCap web application[41,42]. 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 to 9 scale, the importance of items in the initial list proposed by the research team. Ratings of 1 to 3 on the scale were defined as

'not important', 4 to 6 as 'important but not critical' and 7 to 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 (BV and MN) and conflict was resolved by consensus[43]. 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, interquartile range, 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 cut-offs, 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 Open Science Framework (OSF)[29]. All analyses of item scores and comments were performed independently by two members of the research team (BV and MN), using NVivo (QSR International Pty Ltd., v1.0) and Python (Python Software Foundation, v.3.8.5). Conflicts were resolved by consensus.

The initial item list contained 54 items. 120 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, 9 items remained unchanged, 22 were reworded/completed, 21 reorganised (merged/split, becoming 13 items), 2 items dropped, and 9 new items 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 numerous comments describing them as an entirely separate aspect of evaluation. The revised list was reorganised into items and subitems. 136 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, 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 OSF[29].

Consensus meeting

A virtual consensus meeting was held on three occasions between the 14th and the 16th of June 2021, to debate and agree the content and wording of the DECIDE-AI reporting guideline. The 16 members of the Consensus Group (see Suppl. notes 1, Suppl. Table 2a and 2b) 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 cut-off 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 study. The

Consensus Group selected 17 items (made of 28 subitems in total) for inclusion in the AI-specific list and 10 items for inclusion in the generic reporting item list. A summary of the Consensus Group votes can be found in Suppl. table 3.

Qualitative evaluation

The drafts of the guideline and of the Explanation and Elaboration (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 (Suppl. notes 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 OSF[29]). Item wording amendments and modifications to the E&E sections were conducted based on the feedback from the qualitative evaluation, which was independently analysed by two reviewers (BV and MN) 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. Suppl. figures 1 and 2 provide graphical representations of the two item lists' (AI-specific and generic) evolution.

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 (see Figure 1 and Table 1). Depending on the chosen study design and if available, authors may also wish to complete the reporting according to study type specific guideline (e.g. STROBE for cohort studies)[44]. Table 2 presents the DECIDE-AI checklist, comprising of the 17 AI-specific reporting items and 10 generic reporting items selected by the Consensus Group. Each item comes with an E&E to explain why and how reporting is recommended (see 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[45], 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 of professional background and experience. The level of interest across stakeholder groups and the high response rate amongst 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 realising 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 recognised development pathways for complex interventions[1,8,9,46], and aligns conceptually with proposed frameworks for clinical AI[6,47–49], 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 crystallise 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 impact 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. While 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[50].

The Steering Group paid special attention to the integration of DECIDE-AI within the broader scheme of AI guidelines (e.g. TRIPOD-AI, STARD-AI, SPIRIT-AI and CONSORT-AI). It also focussed on DECIDE-AI being applicable to all type of decision support modalities (i.e. detection, diagnostic, prognostic, and therapeutic). The final checklist should be considered as minimum scientific reporting standards and do not preclude reporting additional information, nor are they 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 (e.g. based on the stated intended use within the study), and are part of the clinical evidence generation process (e.g. 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[51], 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.

Discussion

Several topics led to more intense discussion than others, both during the Delphi process and Consensus Group discussion. Regardless of whether the corresponding items were included or not, 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 (e.g. reasons for variation from the AI system recommendation, user satisfaction, etc.), who might logically be consented as study participants, and therefore should be considered as such. Because user characteristics (e.g. 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 (e.g. 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 can often be inferred from other sources, like 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 contextualise the system's outputs within the broader clinical

information environment[19], and potentially for regulatory purpose[52]. 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[53,54]. 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*[55], how much impact 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 towards 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 operationalise 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 optimised, fixed version during later summative evaluation[8,9,56,57].

Sixth, several experts noted the benefit of conducting human factors evaluation prior to clinical implementation and considered that therefore human factors should be reported separately. However, even robust preclinical human factors evaluation will not reliably characterise all the potential human factors issues which 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 only happen in the context of its clinical human factors evaluation.

Finally, several experts raised concerns that the DECIDE-AI guideline prescribes an evaluation 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 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 (e.g. OSF) is recommended to allow the sharing and connecting of all required information within one main published evaluation report.

There are several limitations to our work which should be considered. First, the issue of potential biases, which apply to any consensus process: these include anchoring or participant selection biases[58]. The research team tried to mitigate bias through the survey design, using open-ended questions analysed through a thematic analysis, and via 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 towards 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 amongst European/UK scientists (25 out 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 between 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 we started developing the DECIDE-AI guideline. This might have impacted the exhaustiveness of the initial item list created from literature review. However, the wide range of stakeholders involved and 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 realise 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 lay the foundations for both larger clinical studies and later widespread adoption.

STATEMENTS AND AUTHOR INFORMATION

Data availability

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.

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 three years after publication of this manuscript.

Acknowledgment

The authors would like to thank all Delphi participants and experts who participated in the guidelines qualitative evaluation. BV would also like to thank Benjamin Beddoe (Sheffield Teaching Hospital), Nicole Bilbro (Maimonides Medical Center), Neale Marlow (Oxford University Hospitals), Elliott Taylor (Nuffield Department of Surgical Science, University of Oxford) and Stephan 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. BV is funded by a Berrow Foundation Lord Florey scholarship. MN is supported by the UKRI CDT in AI for Healthcare (http://ai4health.io - grant No. P/S023283/1). DC receives funding from Wellcome Trust, AstraZeneca, RCUK and GlaxoSmithKline. GSC is supported by the NIHR Biomedical Research Centre, Oxford, and Cancer Research UK

(programme grant: C49297/A27294). MI is supported by a Maimonides Medical Center Research fellowship. XL receives funding from the Wellcome Trust, the National Institute of Health Research/NHSX/Health Foundation, the Alan Turing Institute, the MHRA, and NICE. BAM 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. MM receives funding from the Dalla Lana School of Public Health and Leong Centre for Healthy Children. JO is employed by the Medicines and Healthcare products Regulatory Agency, the competent authority responsible for regulating medical devices and medicines within the UK. Elements of the work relating to the regulation of AI as a medical device are funded via grants from NHSX and the Regulators' Pioneer Fund (Department for Business, Energy, and Industrial Strategy). SS receives grants from the National Science Foundation, the American Heart Association, the National Institute of Health, and the Sloan Foundation. DSWT is supported by the National Medical Research Council, Singapore (NMRC/HSRG/0087/2018;MOH-000655-00), National Health Innovation Centre, Singapore (NHIC-COV19-2005017), SingHealth Fund Limited Foundation (SHF/HSR113/2017), Duke-NUS Medical School (Duke-NUS/RSF/2021/0018;05/FY2020/EX/15-A58), and the Agency for Science, Technology and Research (A20H4g2141; H20C6a0032). PWa is supported by the NIHR Biomedical Research Centre, Oxford and holds grants from the NIHR and Wellcome. PMc receives grants from Medtronic (unrestricted educational grant to Oxford University for the IDEAL Collaboration) and 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 guidelines. These views do not necessarily reflect those of their institutions or funders.

Contribution

BV, MN and PMcC designed the study. BV and MI conducted the literature searches. Members of the DECIDE-AI Steering Group (BV, DC, GSC, AKD, LF, BG, XL, PMa, LM, SS, PWa, PMc) provided methodological input and oversaw the conduct of the study. BV and MN conducted the thematic analysis, Delphi rounds analysis, and produced the Delphi round summaries. Members of the DECIDE-AI Consensus Group (BV, GSC, SP, BG, XL, BAM, PM, MM, LM, JO, CR, SS, DSWT, WW, PWh, PMc) selected the final content and wording of the guideline. BC chaired the consensus meeting. BV, MN and BC 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.

Conflict of interest

MN consults for Cera Care, a technology enabled homecare provider. BC was a Non-Executive Director of the UK Medicines and Healthcare products Regulatory Agency (MHRA) from September 2015 until 31 August 2021. DC receives consulting fees from Oxford University Innovation, Biobeats, Sensyne Health, and has advisory role with Bristol Myers Squibb. BG 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. XL 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). PMa is the co-founder of BrainX LLC and BrainX Community LLC. MM reports consulting fees

from AMS Healthcare, and honoraria from the Osgoode Law School and Toronto Pain Institute. LM is director and owner of Morgan Human Systems. JO holds an honorary post as an Associate of Hughes Hall, University of Cambridge. CR is an employee of HeartFlow Inc., including salary and equity. SS has received honoraria from several universities and pharmaceutical companies for talks on digital health and AI. SS has advisory roles in Child Health Imprints, Duality Tech, Halcyon Health, and Bayesian Health. SS 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. DSWT holds patents linked to AI driven technologies, and a co-founder and equity holder for EyRIS Pte Ltd. PWa declares grants, consulting fees and stocks from Sensyne Health and holds patents linked to AI driven technologies. PMc has advisory role for WEISS International and the technology incubator PhD programme at University College London. BV, GSC, AKD, LF, MI, BAM, SD, PWh, and WW have no further conflict of interest to declare.

Consortium

The DECIDE-AI expert group

Aaron Y. Lee[32]; Alan G. Fraser[33]; Alastair K. Denniston[12,13,14]; Ali Connell[34]; Alykhan Vira[35]; Andre Esteva[36]; Andrew D. Althouse[37]; Andrew L. Beam[38]; Anne de Hond[39]; Anne-Laure Boulesteix[40]; Anthony Bradlow[41]; Ari Ercole[42]; Arsenio Paez[43]; Athanasios Tsanas[44]; Baptiste Vasey[1,2,3]; Barry Kirby[45]; Bart Geerts[15]; Ben Glocker[46]; Bilal A. Mateen[8,17,18]; Bruce Campbell[5,6]; Campbell Rogers[24]; Carmelo Velardo[2,47]; Chang Min Park[48]; Charisma Hehakaya[49]; Chris Baber[50]; Chris Paton[51]; Christian Johner[52]; Christopher J. Kelly[34]; Christopher J. Vincent[53]; Christopher Yau[54]; Clare McGenity[55]; Constantine Gatsonis[56]; Corinne Faivre-Finn[57]; Crispin Simon[58]; Daniel S.W. Ting[27,28]; Danielle Sent[59]; Danilo Bzdok[60]; Darren Treanor[61]; David A. Clifton[2]; David C. Wong[62]; David F. Steiner[63]; David Higgins[64]; Dawn Benson[65]; Declan P. O'Regan[66]; Dinesh V. Gunasekaran[27]; Dominic Danks[67]; Emanuele Neri[68]; Evangelia Kyrimi[69]; Falk Schwendicke[70]; Farah Magrabi[71]; Frances Ives[72]; Frank E. Rademakers[73]; Gary S. Collins[7]; George E. Fowler[74]; Giuseppe Frau[75]; H. D. Jeffry Hogg[76]; Hani J. Marcus[77]; Heang-Ping Chan[78]; Henry Xiang[79]; Hugh F. McIntyre[80]; Hugh Harvey[81]; Hyungjin Kim[82]; Ibrahim Habli[83]; James C. Fackler[84]; James Shaw[85]; Janet Higham[86]; Jared M. Wohlgemut[87]; Jaron Chong[88]; Jean-Emmanuel Bibault[89]; Jérémie F. Cohen[90]; Jesper Kers[91]; Jessica Morley[92]; Joachim Krois[93]; Joao Monteiro[94]; Joel Horovitz[16]; Johan Ordish[23]; John Fletcher[30]; Jonathan Taylor[95]; Jung Hyun Yoon[96]; Karandeep Singh[97]; Karel G.M. Moons[98]; Kassandra Karpathakis[99]; Ken Catchpole[100]; Kerenza Hood[101]; Konstantinos Balaskas[102]; Konstantinos Kamnitsas[50]; Laura Militello[103]; Laure Wynants[104]; Lauren Morgan[22]; Lauren Oakden-Rayner[105]; Laurence B. Lovat[106]; Livia Faes[14]; Luc J.M. Smits[107]; Ludwig C. Hinske[108]; M. Khair ElZarrad[109,§]; Maarten van Smeden[110]; Mara Giavina-Bianchi[111]; Mark Daley[112]; Mark P. Sendak[113]; Mark Sujan[114];

Maroeska Rovers[115]; Matthew DeCamp[116]; Matthew Woodward[117]; Matthieu Komorowski[118]; Max Marsden[87]; Maxine Mackintosh[119]; Melissa D. McCradden[20,21]; Michael D. Abramoff[120]; Miguel Ángel Armengol de la Hoz[121]; Myura Nagendran[4]; Neale Hambidge[122]; Neil Daly[123]; Niels Peek[124]; Oliver Redfern[125]; Omer F. Ahmad[126]; Patrick M. Bossuyt[127]; Pearse A. Keane[128]; Pedro N.P. Ferreira[129]; Peter McCulloch[1]; Peter Watkinson[3,29]; Peter Wheatstone[31]; Petra Schnell-Inderst[130]; Pietro Mascagni[131]; Piyush Mathur[19]; Prokar Dasgupta[132]; Pujun Guan[133]; Rachel Barnett[15]; Rawen Kader[134]; Reena Chopra[34]; Ritse M. Mann[135]; Rupa Sarkar[136]; Saana M. Mäenpää[137]; Samuel G. Finlayson[138]; Sarah Vollam[3]; Sebastian J. Vollmer[139]; Seong Ho Park[140]; Shakir Laher[141]; Shalmali Joshi[142]; Siri L. van der Meijden[15,143]; Spiros Denaxas[8,9,10,11]; Suchi Saria[25,26]; Susan C. Shelmerdine[144]; Tien-En Tan[27]; Tom J.W. Stocker[145]; Valentina Giannini[146]; Vince I. Madai[147]; Virginia Newcombe[148]; Wei Yan Ng[27]; Wendy A. Rogers[149]; William Ogallo[150]; Wim Weber[30]; Xiaoxuan Liu[12,13]; Yoonyoung Park[151]; Zane B. Perkins[87].

[32] Department of Ophthalmology, School of Medicine, University of Washington, Seattle, WA, USA;
[33] School of Medicine, Cardiff University, Cardiff, UK;
[34] Google Health, London, UK;
[35] Quantium Health, Johannesburg, SA;
[36] Artera Research, Artera, Mountain View, CA, USA;
[37] University of Pittsburgh, Pittsburgh, PA, USA;
[38] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA;
[39] CAIRElab, Leiden University Medical Centre, Leiden, NL;
[40] Institute for Medical Information Processing, Biometry and Epidemiology, Ludwig Maximilian University, Munich, DE;
[41] Rheumatology Department, Royal Berkshire Hospital, Reading, UK;
[42] Cambridge Centre for AI in Medicine, University of Cambridge, Cambridge, UK;
[43] Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK;
[44] Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK;
[45] K Sharp, Llanelli, UK;
[46] Department of Computing, Imperial College London, London, UK;
[47] Sensyne Health, UK;
[48] Seoul National University College of Medicine, Seoul, KR;
[49] Division of Imaging & Oncology, University Medical Center Utrecht, Utrecht, NL;
[50] School of Computer Science, University of Birmingham, Birmingham, UK;
[51] Nuffield Department of Medicine, University of Oxford, Oxford, UK;
[52] Johner Institute, Konstanz, DE;
[53] PDD Group Ltd, London, UK;
[54] University of Manchester, Manchester, UK;
[55] Pathology and Data Analytics, University of Leeds, Leeds, UK;
[56] Department of Biostatistics, Brown University School of Public Health, Providence, RI, USA;
[57] The Christie NHS Foundation Trust, Manchester, UK;

[58] London School of Economics, London, UK;
[59] Department of Medical Informatics, Amsterdam UMC, University of Amsterdam, Amsterdam, NL;
[60] Mila, Quebec Artificial Intelligence Institute, Montreal, CA;
[61] Leeds Teaching Hospitals NHS Trust, Leeds, UK;
[62] Department of Computer Science and Centre for Health Informatics, University of Manchester, Manchester, UK;
[63] Google Health, Palo Alto, USA;
[64] Berlin Institute of Health, Berlin, DE;
[65] Healthcare Safety Investigation Branch, Farnborough, UK
[66] MRC London Institute of Medical Sciences, Imperial College London, London, UK;
[67] Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK;
[68] University of Pisa, Pisa, IT;
[69] School of Electronic Engineering and Computer Science (EECS), Queen Mary University of London, London, UK;
[70] Charité Universitätsmedizin Berlin, Berlin, DE;
[71]Australian Institute of Health Innovation, Macquarie University, Sydney, AU;
[72] West Midlands Academic Health Science Network, Birmingham, UK;
[73] Department Cardiovascular sciences, KU Leuven, Leuven, BE;
[74] Bristol Centre for Surgical Research, Department of Population Health Sciences, Bristol Medical School, Bristol, UK;
[75] Deep Blue, Rome, IT;
[76] Population Health Science Institute, Newcastle University, Newcastle upon Tyne, UK;
[77] Department of Neurosurgery, National Hospital for Neurology and Neurosurgery, Queen Square, London, UK;
[78] Department of Radiology, University of Michigan, Ann Arbor, MI, USA;
[79] The Abigail Wexner Research Institute, Nationwide Children's Hospital, The Ohio State University, Columbus, OH, USA;
[80] Department of Medicine, East Sussex Healthcare Trust, Hastings, UK;
[81] Hardian Health, UK;
[82] Department of Radiology, Seoul National University Hospital, Seoul, KR;
[83] Department of Computer Science, University of York, York, UK;
[84] Department of Anesthesiology and Critical Care Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD, USA;
[85] Joint Centre for Bioethics, University of Toronto, Toronto, CA;
[86] University of Oxford, Oxford, UK;
[87] Centre for Trauma Sciences, Blizard Institute, Queen Mary University of London, London, UK;
[88] Medical Imaging, Western University, London, CA;
[89] Radiation Oncology Department, Hôpital Européen Georges Pompidou, AP-HP, Paris, FR;
[90] Center of Research in Epidemiology and Statistics (Inserm 1153), Université de Paris, Paris, FR;
[91] Department of Pathology, Amsterdam UMC, University of Amsterdam, Amsterdam, NL;
[92] Oxford Internet Institute, University of Oxford, Oxford, UK;
[93] Oral Diagnostics & Digital Health & Health Services Research, Charité Universitätsmedizin Berlin, Berlin, Germany;
[94] Nature Medicine, New York, NY, USA;
[95] Nuclear Medicine / 3DLab, Sheffield Teaching Hospitals, Sheffield, UK;
[96] Department of Radiology, Severance Hospital, Yonsei University College of Medicine, Seoul, KR;
[97] Department of Learning Health Sciences, University of Michigan Medical School, Ann Arbor, MI, USA;
[98] Julius Center, UMC Utrecht, Utrecht University, Utrecht, NL;
[99] Harvard TH Chan School of Public Health, Boston, MA, USA;
[100] Medical University of South Carolina, Charleston, SC, USA;
[101] Centre for Trials Research, Cardiff University, Cardiff, UK;
[102] Moorfields Ophthalmic Reading Centre and Clinical AI Hub, Moorfields Eye Hospital, London, UK;
[103] Applied Decision Science LLC, USA;
[104] Department of Epidemiology, CAPHRI Care and Public Health Research Institute, Maastricht University, Maastricht, NL;
[105] Australian Institute of Machine Learning, University of Adelaide, Adelaide, AU;
[106] University College London, London, UK;
[107] Department of Epidemiology, Maastricht University, Maastricht, NL;

[108] Institute for Biomedical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-University Munich, Munich, DE;
[109] U.S. Food and Drug Administration;
[110] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Utrecht, NL;
[111] Hospital Israelita Albert Einstein, São Paulo, BR;
[112] The University of Western Ontario, London, CA;
[113] Duke Institute for Health Innovation, Durham, USA;
[114] Human Factors Everywhere Ltd, Woking, UK;
[115] Department of Operating rooms, Radboudumc, Nijmegen, NL;
[116] University of Colorado, Boulder, CO, USA;
[117] THIS Institute, School of Clinical Medicine, University of Cambridge, Cambridge, UK;
[118] Dept of Surgery and Cancer, Imperial College London, London, UK;
[119] Genomics England, Queen Mary University of London, London, United Kingdom;
[120] University of Iowa, Iowa City, IA, USA;
[121] Big Data Department, FPS, Regional Ministry of Health of Southern Spain, ES;
[122] National Hospital for Neurology and Neurosurgery, Queen Square, London, UK;
[123] Skin Analytics, London, UK;
[124] Division of Informatics, Imaging and Data Science, The University of Manchester, Manchester, UK;
[125] Kadoorie Centre for Critical Care Research and Education, Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, UK;
[126] Wellcome/EPSRC centre for Interventional & Surgical Sciences (WEISS), University College London, London, UK;
[127] Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, NL;
[128] Institute of Ophthalmology, University College London, London, UK;
[129] CENTEC - IST, University of Lisbon, Lisbon, PT;
[130] Institut of Public Health, Medical Decision Making and HTA, UMIT - University for Health Sciences, Medical Informatics and Technology, Hall i. T., AT;
[131] Gastrointestinal Endoscopic Surgery, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome, IT;
[132] King's Health Partners Academic Surgery, King's College London, London, UK;
[133] Graduate School of Biomedical Sciences, University of Texas MD Anderson Cancer Center and UTHealth, Houston, USA;
[134] Division of Surgery and Interventional Sciences, University College London, London, UK;
[135] Department of medical imaging, Radboud University Medical Center, Nijmegen, NL;
[136] The Lancet Digital Health, The Lancet Group, London, UK;
[137] Department of Neurosurgery, Helsinki University Hospital, Helsinki, FI;
[138] Harvard Medical School, Boston, MA, USA;
[139] Data Science and its Application, DFKI, Kaiserslautern, DE;
[140]Department of Radiology, Asan Medical Center, Seoul, KR;
[141] University of York, York, UK;
[142] SEAS, Harvard University, Cambridge, MA, USA;
[143] Leiden University Medical Center, Leiden, NL;
[144] Department of Clinical Radiology, Great Ormond Street Hospital for Children NHS Foundation Trust, London, UK;
[145] PUBLIC ltd, Oxford, UK;
[146] University of Turin, Turin, IT;
[147] QUEST Centre for Responsible Research, Berlin Institute of Health, Charité Universitätsmedizin Berlin, Berlin, Germany;
[148] University Division of Anaesthesia, Department of Medicine, University of Cambridge, Cambridge, UK;
[149] Philosophy Department and School of Medicine, Macquarie University, Sydney, AU;
[150] IBM Research Africa, Nairobi, KE;
[151] Center for Computational Health, IBM Research, Cambridge, MA USA;
§ Participation represents personal views and perspectives that may not necessarily reflect the positions and opinions of the U.S. FDA.

FIGURES

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

TABLES

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, 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-randomised controlled trials, …)* with additional features such as modification of intervention, analysis of prespecified 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	**Randomised controlled trials (protocol)**	Extension of SPIRIT. Used to report the protocols of randomised controlled trials evaluating AI systems as interventions.
CONSORT-AI	Comparative prospective evaluation	**Randomised controlled trials**	Extension of CONSORT. Used to report randomised 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.

Table 1. Overview of existing and upcoming AI reporting guidelines. The bold font indicates the primary target of the guidelines, either a specific stage or a specific study design. *Although existing reporting guidelines exist for some of these study designs (e.g. STROBE for cohort studies), none of them cover 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 studies.

Item nº	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 artificial intelligence 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, 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 impact, including patient outcomes, 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 familiarise the users with the AI system, including any training received prior to 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, 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, analysed, and minimised.
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 analysed, as well as any prespecified additional analyses, including subgroup analyses and their rationale.
8	Ethics	Describe whether specific methodologies were utilised 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 prespecified outcomes, including outcomes for any comparison group if applicable.
VIII	Subgroups analysis	Report on the differences in the main outcomes according to the prespecified 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.
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 impacts 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 recognised 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.

Table 2. DECIDE-AI checklist. AI-specific items are numbered in Arab numerals, generic items in Roman numerals; AI = artificial intelligence.

BOXES

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 (either due to early prototyping, version updates, or continuous learning design) and to analyse related performance changes.
- minimise the potential of this technology to embed and reproduce existing health inequality and systemic biases.
- estimate the generalisability of findings across sites and populations.
- enable reproducibility of the findings in the context of a dynamic innovation field and intellectual property protection.

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

AI system	Decision support system incorporating AI and consisting of: (i) the artificial intelligence or machine learning algorithm; (ii) the supporting software platform; and (iii) 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.
Artificial intelligence (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."[59]
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, analysing clinical data and using scientific methods, to evaluate the clinical performance, effectiveness and/or safety of an AI system, when used as intended[35].
Clinical investigation	Study performed on one or more human subjects to evaluate the clinical performance, effectiveness and/or safety of an AI system[60]. This can be performed in any setting (e.g. community, primary care, 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- 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	Bidirectional 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, mode of action.
Learning curves	Graphical plotting of user performance against experience[61]. 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 impact on patient care. As opposed to "offline" or "shadow mode" evaluation where the decisions do not have a direct impact 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 which 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. *NB: 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 wellbeing 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"[62].
User	Person interacting with the AI system to inform their decision making. This person could be a healthcare professional or a patient.

Box 2. Glossary of terms. The definitions given 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.

REFERENCES

1. Skivington, K. *et al.* A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ* 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. Heal.* 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, (2021).

4. Freeman, K. *et al.* Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *BMJ* 374, n1872 (2021).

5. A. Keane, P. & J. Topol, E. *With an eye to AI and autonomous diagnosis*. *npj Digital Medicine* vol. 1 (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, (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, 2054270415616548–2054270415616548 (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 Devices - Guidance for Industry and Food and Drug Administration Staff*. (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*. (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 Heal. & Care Informatics* 26, e100081 (2019).

20. Sujan, M., Baber, C., Salmon, P., Pool, R. & Chozos, N. *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. *BMJ* 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. Moher, D., Schulz, K. F., Simera, I. & Altman, D. G. Guidance for Developers of Health Research Reporting Guidelines. *PLOS Med.* 7, e1000217 (2010).

28. Dalkey, N. & Helmer, O. An Experimental Application of the DELPHI Method to the Use of Experts. *Manage. Sci.* 9, 458–467 (1963).

29. Vasey, B., Nagendran, M. & McCulloch, P. Open Science Framework - DECIDE-AI reporting guideline. (2021). https://osf.io/qfy73/?view_only=027338bc2bfc4f14ba530b049faac81c.

30. Vollmer, S. *et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* 368, l6927 (2020).

31. 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, (2021).

32. 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* (2019) doi:10.1007/s11948-019-00165-5.

33. Xie, Y. *et al.* Health Economic and Safety Considerations for Artificial Intelligence Applications in Diabetic Retinopathy Screening. *Transl. Vis. Sci. Technol.* 9, 22 (2020).

34. Norgeot, B. *et al.* Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.* 26, 1320–1324 (2020).

35. IMDRF Medical Device Clinical Evaluation Working Group. *Clinical Evaluation*. (2019).

36. IMDRF Software as Medical Device (SaMD) Working Group. *'Software as a Medical Device': Possible Framework for Risk Categorization and Corresponding Considerations*. (2014).

37. National Institute for Health and Care Excellence (NICE). *Evidence standards framework for digital health technologies*. (2019).

38. Accelerated Access Collaborative & NHSx. *AI-Award Evaluation Playbook - Version 1*. (2020).

39. High-Level Independent Group on Artificial Intelligence (AI HLEG). *Ethics Guidelines for Trustworthy AI*. *European Commission* vol. 32 https://ec.europa.eu/digital (2019).

40. 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).

41. 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).

42. Harris, P. A. *et al.* The REDCap consortium: Building an international community of software platform partners. *J. Biomed. Inform.* 95, 103208 (2019).

43. 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).

44. von Elm, E. *et al.* Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 335, 806–808 (2007).

45. Page, M. J. *et al.* The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, n71 (2021).

46. Sedrakyan, A. *et al.* IDEAL-D: A rational framework for evaluating and regulating the use of medical devices. *BMJ* 353, i2372 (2016).

47. Park, Y. *et al.* Evaluating artificial intelligence in medicine: phases of clinical research. *JAMIA Open* 3, 326–331 (2020).

48. 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).

49. Sendak, M. P. *et al.* A path for translation of machine learning products into healthcare delivery. *EMJ Innov.* (2020) doi:DOI/10.33590/emjinnov/19-00172.

50. Moher, D., Jones, A., Lepage, L. & Group, for the C. Use of the CONSORT Statement and Quality of Reports of Randomized TrialsA Comparative Before-and-After Evaluation. *JAMA* 285, 1992–1995 (2001).

51. Park, S. H. Regulatory Approval versus Clinical Validation of Artificial Intelligence Diagnostic Tools. *Radiology* 288, 910–911 (2018).

52. 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).

53. Lipton, Z. C. The Mythos of Model Interpretability. *Commun. ACM* 61, 36–43 (2018).

54. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit. Heal.* 3, e745–e750 (2021).

55. 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).

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

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

58. Winkler, J. & Moser, R. Biases in future-oriented Delphi studies: A cognitive perspective. *Technol. Forecast. Soc. Change* 105, 63–76 (2016).

59. International Organization for Standardization. *Information technology - Artificial intelligence (AI) - Bias in AI systems and AI aided decision making (ISO/IEC TR 24027:2021)*. (2021).

60. IMDRF Medical Device Clinical Evaluation Working Group. *Clinical Investigation*. (2019).

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

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