ScienceDirect

Contents lists available at sciencedirect.com
Journal homepage: www.elsevier.com/locate/jval

Methodology

Consolidated Health Economic Evaluation Reporting Standards for Interventions That Use Artificial Intelligence (CHEERS-AI)

Jamie Elvidge, MSc, Claire Hawksworth, MSc, Tuba Saygın Avşar, PhD, Antal Zemplenyi, PhD, Anastasia Chalkidou, PhD, Stavros Petrou, PhD, Zsuzsanna Petykó, PhD, Divya Srivastava, PhD, Gunjan Chandra, MSc, Julien Delaye, MA, Alastair Denniston, PhD, Manuel Gomes, PhD, Saskia Knies, PhD, Petros Nousios, MSc, Pekka Siirtola, PhD, Junfeng Wang, PhD, Dalia Dawoud, PhD, on behalf of the The CHEERS-AI Steering Group

ABSTRACT

Objectives: Economic evaluations (EEs) are commonly used by decision makers to understand the value of health interventions. The Consolidated Health Economic Evaluation Reporting Standards (CHEERS 2022) provide reporting guidelines for EEs. Healthcare systems will increasingly see new interventions that use artificial intelligence (AI) to perform their function. We developed Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI (CHEERS-AI) to ensure EEs of AI-based health interventions are reported in a transparent and reproducible manner.

Methods: Potential CHEERS-AI reporting items were informed by 2 published systematic literature reviews of EEs and a contemporary update. A Delphi study was conducted using 3 survey rounds to elicit multidisciplinary expert views on 26 potential items, through a 9-point Likert rating scale and qualitative comments. An online consensus meeting was held to finalize outstanding reporting items. A digital health patient group reviewed the final checklist from a patient perspective.

Results: A total of 58 participants responded to survey round 1, 42, and 31 of whom responded to rounds 2 and 3, respectively. Nine participants joined the consensus meeting. Ultimately, 38 reporting items were included in CHEERS-AI. They comprised the 28 original CHEERS 2022 items, plus 10 new AI-specific reporting items. Additionally, 8 of the original CHEERS 2022 items were elaborated on to ensure AI-specific nuance is reported.

Conclusions: CHEERS-AI should be used when reporting an EE of an intervention that uses AI to perform its function. CHEERS-AI will help decision makers and reviewers to understand important AI-specific details of an intervention, and any implications for the EE methods used and cost-effectiveness conclusions.

Keywords: artificial intelligence, CHEERS checklist, health economic evaluation, reporting guideline.

VALUE HEALTH. 2024; ■(■):■–■

Highlights

- The use of artificial intelligence (AI) in healthcare is expanding rapidly. New health interventions that use AI to perform their functions are increasingly expected to be developed. To date, the reporting of economic evaluations (EEs) of AI-based health interventions appears to lack important details regarding the AI nature of the intervention and potential implications for cost-effectiveness results.
- The Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI (CHEERS-AI) checklist is intended to standardize reporting of EEs of health technologies that use AI. Developed using a Delphi study, it contains 38 reporting items in total. It comprises the original 28 CHEERS-2022 checklist items with 8 elaborations to draw out potential AI-related nuances, plus 10 new AI-specific items that extend upon CHEERS-2022.
- The CHEERS-AI checklist will ensure that important details relating to the AI nature of the intervention and implications for the analysis are reported in a transparent and reproducible way. CHEERS-AI will also support the interpretation and comparison of such studies by reviewers and decision makers. It will raise the standard of EEs reporting for AI technologies as their presence in healthcare proliferates.

Introduction

Healthcare decision makers, health technology assessment (HTA) agencies and reimbursement and pricing authorities often examine the cost-effectiveness of health interventions, alongside clinical effectiveness and other relevant considerations. Cost-effectiveness estimates are typically informed by an economic evaluation (EE), conducted alongside a clinical study or using a decision model, and indicate the likely economic value of an intervention compared with the most appropriate comparator(s). The Consolidated Health Economic Evaluation Reporting Standards (CHEERS 2022) provide researchers with a framework for reporting EEs.[1] By reporting their EE using the CHEERS 2022 checklist, authors should provide at least a minimum standard of transparency and reproducibility.

The use of artificial intelligence (AI) in healthcare is an area of rapid development, with an increasing number of AI-enabled healthcare technologies reaching late stage development and achieving market authorization as medical devices.[2] AI essentially describes the use of computer algorithms

that perform tasks normally associated with human intelligence.[3] Decision-making organizations are likely to increasingly be presented with health interventions that use AI systems, such as recent examples to analyze chest X-rays[4] and assist radiotherapy treatment planning.[5] This is reflected by the National Institute for Health and Care Excellence recently updating its framework for digital health technologies to include evidence standards for AI-based technologies.[3] Given that AI-driven capabilities are often complex and may affect the cost-effectiveness of the intervention, it is important that the methods used to evaluate AI-based interventions are clearly reported.

A systematic review of EEs of AI-based interventions,[6] and a recent update of the review,[7] both concluded that advancements in AI appear to be accelerating quicker than the practices of EE. There appears to be notable reporting gaps in published EEs, such as how the AI intervention works, including its role in and influence on care pathways. These are likely to be different to commonly evaluated health technologies, such as medicines. We considered that an extension to CHEERS 2022 for the evaluation of AI-based interventions may be warranted to ensure that key information about the AI component(s) and implications for the EE are adequately reported.

This study reports the development of the Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI (CHEERS-AI) reporting guideline extension. CHEERS-AI is intended to extend the original CHEERS statement, by drawing out important nuances relating to AI that might otherwise be missed and introducing new AI-specific reporting items in situations which are appropriate. The key intended audience for CHEERS-AI is likely to be health economists who are reporting an EE of an AI-based intervention. Additionally, healthcare decision makers (and the HTA agencies that support them), journal editors, and peer reviewers may find that CHEERS-AI provides useful AI-specific review criteria when assessing the suitability of a published EE.

This study has been supported by Next-Generation HTA (HTx), which is a Horizon 2020 project supported by the European Union, lasting for 5 years from January 2019 onward.

Methods

The development of CHEERS-AI followed the methodology used to develop CHEERS 2022[1] and recent AI extensions to standards for reporting clinical trial protocols[8] and results[9] and the Enhancing the Quality and Transparency of Health Research guidance.[10,11] The study record was published on the Enhancing the Quality and Transparency of Health Research website,[12] and the study protocol was published in full elsewhere.[13] Ethical approval was granted by the Newcastle University Research Ethics Committee (28568/2022).

Our methodology consisted of 3 phases with multiple steps. The initiation phase involved (1) convening an expert steering group, (2) creating a longlist of potential reporting items, and (3) agreeing on a shorter list of potential items. The development phase involved 4) a 3-round consensus-building exercise using the Delphi method,[14] (5) a consensus meeting, and (6) engagement with a patient expert group. The finalization phase involved (7) piloting the CHEERS-AI reporting items and (8) final steering group ratification of all items.

Initiation Phase

Steering group

A multidisciplinary steering group (n = 17) was convened to steward the development process and provide expert oversight. The steering group included expertise in health economics (n = 9), HTA (n = 8), AI (n = 5), clinical practice (n = 2), patient advocacy (n = 1), and a member of the CHEERS 2022 Task Force (see Appendix Table 1 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2024.05.006). A prespecified minimum of 50% of steering group members was required for quorate decision making.

Identifying candidate items

Two recent systematic reviews of EEs of AI-based health interventions and their 41 constituent studies were reviewed[6,7] by 4 steering group members (A.Z., C.H., J.E., and Z.P.). The reviewers generated a longlist of potential AI-specific reporting items based on their interpretation of the published EEs. This longlist was reviewed and discussed by the full steering group to reach consensus about those that should proceed to the Delphi study. If there was disagreement about a potential item, the group included the item in the Delphi study.

Development Phase

Delphi study

The steering group nominated relevant stakeholders to invite to participate in the Delphi study based on their expertise and professional networks. Stakeholders included methodologists (A.I. and E.E.), healthcare decision makers (HTA and payers), healthcare professionals, journal editors, research funders, patient advocates, ethicists, and representatives of the pharmaceutical and health technology industries. The 119 identified stakeholders were invited by email to participate in a three-round survey process and encouraged to share the invitation with their networks. The Delphi method was administered between May 2023 and September 2023. The response time was 3 weeks for survey rounds 1 and 2, and 2 weeks for survey round 3. Two reminders were sent before each round closed.

Participants were asked to score the importance of each candidate item on a 9-point Likert scale, from 1 (not important) to 9 (very important). For each item, respondents could provide qualitative comments about the proposed text and description, such as suggestions to improve the wording or merge items. After each survey round, the steering group convened and discussed the quantitative and qualitative findings. The group implemented minor wording changes, and more substantial changes such as merging items, in which they were considered sensible without altering the intended meaning of the proposed items.

Consistent with the CHEERS 2022 Delphi study,[1,11,15,16] items with 70% of expert ratings below 7 in survey round 1 were excluded. Items with a mean score of 7 or higher after round 2 were included, and items with a mean score of 4 or less were excluded. Remaining items, with mean scores higher than 4 but less than 7, proceeded to round 3. After round 3, the steering group would include items that scored 5 or higher, except any that received multiple conflicting qualitative comments. Any such items would be discussed during a consensus meeting with a subgroup of Delphi participants. The steering group would make the final decision about them.

Consensus meeting

A subgroup of Delphi study participants reflecting the various stakeholder groups and countries was invited to join a consensus meeting on 2 October 2023. The purpose was to discuss the results of survey round 3, including outstanding contentious items, and ratify all items that had been accepted after round 2. Minor wording changes to reporting items were implemented provided they received the approval of at least 50% of the Delphi study participants in attendance.

Patient involvement

To ensure we had sufficient patient involvement in the development of CHEERS-AI, the reporting items were shared with the Digital Advisory Group of patient advocates at the European Organisation for Rare Diseases (EURORDIS) to elicit a patient perspective about their presentation and clarity.

Finalization Phase

Pilot exercise

The CHEERS-AI reporting guideline extension was piloted by 5 members of the steering group (A.C., A.Z., D.S., S.P., and T.S.A.) on a sample of 9 published EEs of AI-based interventions identified across 2 systematic reviews (6 and 7). The purpose of the pilot exercise was to identify any unclear, ambiguous, or challenging items when applying CHEERS-AI in practice. Recommended changes would be implemented if they received approval by the steering group.

Steering group ratification

The steering group convened on 19 October 2023 to discuss CHEERS-AI items after the Delphi study and consensus meeting and continued to review any changes in response to the patient group's review and pilot exercise until December 2023. The steering group collectively made the final decision about any changes and ratified the final CHEERS-AI reporting guideline extension.

Results

Initiation Phase

The long list of potential reporting items contained 30 items. The steering group agreed that 26 items were suitable for inclusion in the Delphi study. They were classified as either elaborations on a pre-existing CHEERS 2022 reporting item or extensions to CHEERS 2022, as follows:

- An elaboration adds AI-specific context to a 'parent' CHEERS 2022 reporting item. It indicates that the existing parent item may be sufficient for EEs of AI interventions if the researcher is diligent, but important AI-related nuances need to be pointed out to ensure they are always reported.
- An extension is an entirely new reporting item that goes beyond elaborating on an existing CHEERS 2022 item. It indicates important AI-specific information that must be reported when evaluating AI-based health interventions to ensure transparency and replicability.

Development Phase

In the Delphi study, a total of 58 individuals participated in the first survey, with health economists (53%, 31/58), HTA professionals (26%, 15/58), and AI experts (17%, 10/58) being the best represented stakeholders. The second survey round generated responses from 42 participants (72% retention, 42/58), of whom 31 also responded to the third survey round (74% retention, 31/42). Participants were based in 17 countries across Europe, North America, Asia, and Oceania. Aggregated information about participants' expertise and location in each survey round is provided in the Appendix Table 2 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2024.05.006.

Of the 26 potential reporting items in survey round 1, none were excluded but several were merged to avoid repetition, meaning that 18 potential items proceeded to round 2. Thirteen items met the criteria for inclusion after round 2. Five items proceeded to round 3, of which 1 remained contentious following the survey and was discussed extensively at the subsequent consensus meeting.

A subgroup of 9 Delphi study participants joined the consensus meeting after survey round 3. Seven nonvoting steering group members also attended the meeting to facilitate and contribute to the discussion. One major change was recommended by participant to exclude an extension item because an existing CHEERS 2022 item (#16) was considered to be sufficient. However, this recommendation was in conflict with the quantitative results from round 3. Therefore, the item proceeded for further steering group deliberation to find a resolution. As a compromise, the item was reclassified as an elaboration on CHEERS 2022 item 16 and reworded to improve its specificity.

The EURORDIS Digital Advisory Group of 13 patient advocates provided positive feedback about the completeness and convenience of the CHEERS-AI reporting items. Consequently, 2 items were subject to minor wording changes to improve their clarity without altering the intended meaning.

Finalization Phase

For the pilot exercise, to test using CHEERS-AI in practice, 9 published EE studies[17–25] were selected at random. In general, the checklist was considered to be easy to use and understand. However, for 5 CHEERS-AI items, none of the EEs included in the pilot were found to adequately report them (see Appendix Tables 3 and 4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2024.05.006). This was considered to indicate reporting limitations in the evidence base, rather than a limitations of CHEERS-AI. Therefore, no substantive changes were made following the pilot exercise.

CHEERS-AI Reporting Standards

Table 1 presents the CHEERS-AI reporting items. Each item has a title, number, and description explaining what the author of an EE should report. AI-specific elaborations on a pre-existing CHEERS 2022 item are presented to the right of the corresponding item. New AI-specific items that might not otherwise be captured by CHEERS 2022 are included in their own rows and are denoted by "AI" item numbers (AI 1, AI 2, and so on). For each item, the authors of an EE are expected to indicate the manuscript section, paragraph, or line that contains the required information.

For AI-specific elaboration and extension items, further guidance for reporting is provided in Tables 2 and 3, respectively, including additional information—such as the rationale for the item and definitions of key terms used in it—that may help users to complete the checklist. Examples of suitable reporting of CHEERS-AI items, identified during our pilot exercise, are provided as Appendix Tables 3 and 4 in Supplemental Materials found at https://doi.org/10.1016/j.jval.2024.05.006.

Discussion

Standardized reporting of the use of AI in healthcare is needed as HTA and decision-making organizations are increasingly presented with AI-enabled interventions. There are recent advances in this area for development (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis-Artificial Intelligence [TRIPOD-AI]) and validation (Prediction model Risk Of Bias ASsessment Tool-Artificial Intelligence [PROBAST-AI]) studies,[26] trial protocols (Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence [SPIRIT-AI][8]), and trial reports (Consolidated Standards of

Table 1. Items included in CHEERS-AI reporting guideline extension.

Section/topic	No.	Guidance for reporting	AI elaboration	Reported in section
Title				
Title	1	Identify the study as an economic evaluation and specify the interventions being compared.	Indicate that the intervention involves an AI component that is under evaluation.	-
Abstract				
Abstract	2	Provide a structured summary that highlights context, key methods, results, and alternative analysis.	Specify the purpose of the intervention with an AI component, and the AI technique used.	-
Introduction				
Background and objectives	3	Give the context for the study, the study question, and its practical relevance for decision making in policy or practice.		-
Methods				
Health economic analysis plan	4	Indicate whether a health economic analysis plan was developed and where available.		-
Study population	5	Describe characteristics of the study population (such as age range, demographics, socioeconomic, or clinical characteristics).		-
Setting and location	6	Provide relevant contextual information that may influence findings.		-
Comparators	7	Describe the interventions or strategies being compared and why chosen.	Describe key details of the AI component of the intervention (and comparators, if appropriate), including: a) the classification by intended purpose and risk tier (for digital health technologies); b) the AI technique used; c) whether it is "locked" (static) or adaptive; d) the version under evaluation; e) the purpose of the intervention, including its potential impact on care; f) the intended user(s), and how users interact with it; g) additional requirements to use it; h) how it is expected to provide benefit over the standard of care.	-
User autonomy	AI 1	Indicate whether the AI intervention (and comparators, if appropriate) is directive, or whether the user(s) retains autonomy to make the care decision.		-
Perspective	8	State the perspective(s) adopted by the study and why chosen.		-
Time horizon	9	State the time horizon for the study and why appropriate.		-
Discount rate	10	Report the discount rate(s) and reason chosen.		-
Selection of outcomes	11	Describe what outcomes were used as the measure(s) of benefit(s) and harm(s).	Describe whether the measure(s) chosen to indicate the benefits and harms of the AI intervention (and comparators) relates to health outcomes, diagnostic outcomes, process outcomes, or other/multiple outcomes.	-

continued on next page

Table 1. Continued

Section/topic	No.	Guidance for reporting	AI elaboration	Reported in section
Measurement of outcomes	12	Describe how outcomes used to capture benefit(s) and harm(s) were measured.	For model-based analysis, describe any assumptions used to inform the potential benefit(s) and harm(s) of the AI intervention in the model (and comparators, if appropriate). Describe the plausibility of analyst assumptions, citing any supportive evidence.	-
Measurement of AI effect	AI 2	Describe the data sources (assessment studies) for the AI intervention's impact on outcomes.		-
Measurement of AI learning over time	AI 3	If the AI intervention (and comparators, if appropriate) learns over time, explain how this affects its performance at the individual level and how this was measured.		-
Development of AI component	AI 4	Describe how the AI component of the intervention (and comparators, if appropriate) was developed, including the training data used and how errors and biases were identified, or cite a source that provides this information.		-
Validation of AI component	AI 5	Describe how the AI component of the intervention (and comparators, as appropriate) and its performance estimates were validated, or cite a source that provides this information.		-
Health benefit	AI 6	Describe how the AI intervention (and comparators, if appropriate) could directly or indirectly provide a health benefit.		-
Population differences	AI 7	Describe important differences between the data sources (assessment studies) for the AI intervention's impact on outcomes and the data set that was used to develop the AI intervention (training data set).		-
Valuation of outcomes	13	Describe the population and methods used to measure and value outcomes.		-
Measurement and valuation of resources and costs	14	Describe how costs were valued.	Describe the purchase cost of the AI intervention (and comparators, if appropriate) and what it is composed of. Describe any additional implementation and maintenance costs.	-
Currency, price date, and conversion	15	Report the dates of the estimated resource quantities and unit costs, plus the currency and year of conversion.		-
Rationale and description of model	16	If modeling is used, describe in detail and why used. Report if the model is publicly available and where it can be accessed.	Describe if the AI component of the intervention has influenced the choice of health economic model and explain why.	-
Analytics and assumptions	17	Describe any methods for analyzing or statistically transforming data, any extrapolation methods, and approaches for validating any model used.		-
Modeling of AI learning over time	AI 8	If the AI intervention (and comparators, if appropriate) learns over time at the individual level, describe any assumptions used to model how this learning affects its performance over time.		-
Characterizing heterogeneity	18	Describe any methods used for estimating how the results of the study vary for subgroups.		-

continued on next page

Table 1. Continued

Section/topic	No.	Guidance for reporting	AI elaboration	Reported in section
Characterizing distributional effects	19	Describe how impacts are distributed across different individuals or adjustments made to reflect priority populations.		-
Characterizing uncertainty	20	Describe methods to characterize any sources of uncertainty in the analysis.		-
Approach to engagement with patients and others affected by the study	21	Describe any approaches to engage patients or service recipients, the general public, communities, or stakeholders (such as clinicians or payers) in the design of the study.		-
Results				
Study parameters	22	Report all analytic inputs (such as values, ranges, and references) including uncertainty or distributional assumptions.		-
Summary of main results	23	Report the mean values for the main categories of costs and outcomes of interest and summarize them in the most appropriate overall measure.		-
Effect of uncertainty	24	Describe how uncertainty about analytic judgments, inputs, or projections affect findings. Report the effect of choice of discount rate and time horizon, if applicable.		-
Impact of AI uncertainty	AI 9	Indicate the extent to which features of the AI intervention may contribute to increased uncertainty about its cost-effectiveness.		-
Effect of engagement with patients and others affected by the study	25	Report on any difference patient/service recipient, general public, community, or stakeholder involvement made to the approach or findings of the study.		-
Discussion				
Study findings, limitations, generalizability, and current knowledge	26	Report key findings, limitations, ethical or equity considerations not captured, and how these could affect patients, policy, or practice.	Comment on potential biases associated with the AI intervention (eg, algorithmic bias) and implications for the generalizability and interpretation of results (eg, reinforcing existing health inequalities).	-
Implementation of AI	AI 10	Comment on any requirements needed to integrate the AI intervention (and comparators, as appropriate) into practice, and other implementation considerations relating to the AI component of the intervention, including implications for the interpretation of cost-effectiveness results.		-
Source of funding	27	Describe how the study was funded and any role of the funder in the identification, design, conduct, and reporting of the analysis.		-
Conflicts of interest	28	Report authors conflicts of interest according to journal or International Committee of Medical Journal Editors requirements.		-

Note. Pre-existing CHEERS 2022 items are numbered 1 through 28. Elaborations that add AI-specific context are shown to the right of the corresponding CHEERS 2022 item. New, AI-specific extension items are numbered "AI 1-10." AI indicates artificial intelligence; CHEERS- AI Consolidated Health Economic Evaluation Reporting Standards for Interventions that use AI.

Table 2. AI elaborations on CHEERS 2022: Further guidance for reporting.

Section/topic	No.	Further guidance for reporting
Title		
Title	1	AI refers to algorithmic techniques that analyze large amounts of data for correlations and patterns and use these patterns to simulate the problem-solving and decision-making capabilities of the human mind. This does not include more traditional statistical techniques. If an intervention under evaluation uses AI to perform its function (eg, through algorithms embedded in a digital health technology), then an appropriate term such as "artificial intelligence" should be included in the study title.
Abstract		
Abstract	2	The purpose of the intervention with an AI component (eg, to treat or diagnose, drive clinical management, inform clinical management) and the way the AI works (eg, deep neural networks for image processing and not traditional statistical techniques) should be reported in the abstract.
Methods		
Comparators	7	(a) The classification and risk tier of the AI-based intervention can relate to existing regulatory frameworks, such as the SaMD framework proposed by the IMDRF, and the evidence standards framework for digital health technologies proposed by the NICE. • The SaMD framework classifies digital health technologies by their intended purpose (ie, for critical, serious or nonserious situations) and significance on the healthcare decision (ie, treat or diagnose, drive clinical management, or inform clinical management). • The evidence standards framework for digital health technologies defines tiers based on the potential risk to service users or the system (ie, tier A, no direct patient, health or care outcomes; tier B, interventions to assist peoples' personal health and wellbeing; tier C, interventions for treating, diagnosing, or guiding care choices). (b) The way the AI works (eg, deep neural networks for image processing) should be reported. (c and d) An AI component that is "locked" or static does not change over time, whereas an adaptive AI may continue to learn from data and change over time, potentially affecting outcomes (this may be captured in the evaluation, or the evaluation may focus on a static version of the intervention at a point in time). (e) The purpose of the intervention should be reported. (f) Intended user(s) could include patients, clinicians, healthcare providers, or other agents. (g) Additional requirements could include specific consent processes or staff training. (h) Potential benefits could include clinical effects on health outcomes and economic effects on resource use or system efficiency.
Selection of outcomes	11	An outcome measure is used to quantify the extent to which an intervention has an effect. For example, a study may measure effectiveness (benefits and harms) in terms of changes in health outcomes, diagnostic outcomes (such as improved accuracy), process outcomes (eg, faster decision making), or several outcomes simultaneously.
Measurement of outcomes	12	Assumptions regarding the effect of the AI intervention, such as the use of arbitrary or exploratory input values, should be transparently reported. Their theoretical or scientific basis should be explained.
Measurement and valuation of resources and costs	14	The purchase cost of an intervention with an AI component may include a purchase price and other components, such as the developer implementing the AI into practice or maintaining it over time. There may be other relevant costs to the healthcare system relating to implementation of an AI intervention.
Rationale and description of model	16	A model may be used in a health economic evaluation to estimate the cost-effectiveness of an intervention. Explain if a particular model structure or programming approach, such as individual patient simulation, has been chosen to characterize the AI intervention appropriately.
Discussion		
Study findings, limitations, generalizability, and current knowledge	26	There may be ethical and equity issues associated with AI that are relevant for decision makers alongside the cost-effectiveness results. Biases may include, for example, the AI intervention being developed using a training data set that is not representative of the population of interest.

AI indicates artificial intelligence; IMDRF, International Medical Device Regulators Forum; NICE, National Institute for Health and Care Excellence; SaMD, Software as a Medical Device.

Reporting Trials–Artificial Intelligence [CONSORT-AI][9]). Those checklists should enhance the reporting of the development and assessment of AI health interventions, and now we have developed the CHEERS-AI reporting guideline extension to do the same for EEs. Our reporting standards were developed using a very similar methodological approach to CHEERS 2022, including a Delphi study to identify multidisciplinary consensus, stewarded by an expert steering group. Researchers conducting EEs of AI-based health interventions can use CHEERS-AI to ensure that AI-specific nuances and implications are reported or cited in a transparent and reproducible way.

The importance of an AI extension to CHEERS 2022 was highlighted during the development process, including through qualitative responses during the Delphi study. Respondents noted as follows: first, unlike medicines, AI interventions often involve diagnostic or clinical decision support tools, which require that researchers think carefully about the mechanism by which the intervention affects care pathways, costs, and health outcomes

Table 3. AI extensions to CHEERS 2022: further guidance for reporting.

Section/topic	No.	Further guidance for reporting
Methods		
User autonomy	AI 1	How directly the intervention affects clinical care, and the extent of user autonomy from its outputs, may be defined in the context of existing regulatory frameworks (eg, SaMD): • "Leads to direct care action" could include the intervention being used to definitive diagnosis, or itself being a treatment. • "Drives clinical management" means the intervention aids treatment, diagnosis or decision making in a supportive way. • "Informs clinical management" means no direct care action is used, for example informing users of options or providing information.
Measurement of AI effect	AI 2	This item relates to the evidence informing the impact of the AI intervention. For interventions that directly affect care (eg, treatments), these may be clinical trials to evaluate efficacy. For interventions that drive or inform care (eg, diagnostic algorithms), these may be diagnostic studies that report predictive accuracy. In the absence of evidence, analyst assumptions may be required. This is not the same as "training data," which an AI component might learn from during development.
Measurement of AI learning over time	AI 3	An intervention with a "learning" (adaptive) AI component may become more effective over time as it learns from data collected during its use. How this learning effect was measured should be reported or signposted using a suitable citation.
Development of AI component	AI 4	How the AI component of an intervention was developed should be described or signposted using a suitable citation (eg, a completed TRIPOD-AI checklist for prediction model development).
Validation of AI component	AI 5	Similar to traditional statistical techniques, the internal and external validity of the development of the AI intervention (including, for example, its predictive model) should be described, or signposted using a suitable citation (eg, a completed TRIPOD-AI checklist for prediction model validation).
Health benefit	AI 6	"Health benefit" refers to the way an intervention affects health outcomes, which can be quantified to estimate incremental cost-effectiveness. AI interventions may have different mechanisms by which they are expected to generate health benefits for the people who use them. For example, an AI-based digital health technology designed to inform clinical management (ie, with no direct care action) may have an indirect effect on health outcomes.
Population differences	AI 7	The data used to develop the intervention's AI component may be referred to as its "training" data set. This may differ from the study population that was used to examine the beneficial impact of the intervention compared with alternative options (eg, clinical trials to evaluate efficacy or diagnostic studies to evaluate predictive accuracy).
Modeling of AI learning over time	AI 8	An intervention with a "learning" (adaptive) AI component may become more effective or accurate over time as it learns from data collected during its use. Any modeling assumptions to capture this should be transparently reported.
Results		
Impact of AI uncertainty	AI 9	Uncertainty is usually characterized as random error, parameter uncertainty or structural uncertainty. The AI nature of an intervention may contribute disproportionately to one or more of these types of uncertainty. AI-specific uncertainties and their potential implications on study results should be reported.
Discussion		
Implementation of AI	AI 10	Requirements for implementation may include, for example, necessary software that may be distinct from standard clinical equipment, or a new data acquisition process. Barriers to implementing a new technology may be relevant for decision makers alongside the cost-effectiveness results.

AI indicates artificial intelligence; CHEERS, Consolidated Health Economic Evaluation Reporting Standards; SaMD, Software as a Medical Device; TRIPOD-AI, Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis-Artificial Intelligence.

(item 7 elaboration, items AI 1 and AI 6). Some participants suggested that several CHEERS-AI items could apply more broadly to EEs of diagnostic technologies with or without AI, perhaps reflecting that AI in healthcare is currently dominated by diagnostic interventions, rather than therapeutics. Second, participants explained that CHEERS-AI can help to clarify costs associated with AI-based interventions (item 14 elaboration), clearly indicating when they are exploratory (eg, Van Leeuwen et al[24] and Gomez et al[27]) or not captured (eg, Szymanski et al[28]) because it is unlikely that predictive algorithms underpinning such interventions will be free to use in practice. The use of such interventions may incur one-off or recurring acquisition costs and significant implementation costs for healthcare providers if they cause disruption to care pathways. Furthermore, the potential for AI health interventions to learn from data over time and become more effective has not been observed to date, and our pilot exercise—conducted to stress test the CHEERS-AI checklist in the way it might be applied by reviewers—identified no EEs that reported on this. Indeed, a total of 5 items were found to be inadequately or not reported by EEs in our pilot exercise, and we believe this further demonstrates the need for CHEERS-AI to improve the reporting of such studies. In the context of AI interventions learning over time, items AI 3 ("Measurement of AI learning over time") and AI 8 ("Modeling of AI learning over time") are likely to be enabling CHEERS-AI to accommodate future advances in AI health technologies. In the meantime, version control, including how often adaptive algorithms are updated and whether those updates are included in the purchase cost of the AI intervention, remains an important reporting consideration.

The development of CHEERS-AI had a few noteworthy limitations. First, similar to any consensus-generation process, it was limited by the expertise of the participants involved. Our Delphi group (n = 58) was smaller than some Delphi studies of AI reporting, such as CONSORT-AI (n = 103),[9] but bigger than others, including the comparable extension to CHEERS 2022 for value of information analysis (CHEERS-VOI) (n = 30).[29] Our steering group

(n = 17) was also relatively large (vs CHEERS 2022, n = 15[1] and CHEERS-VOI, n = 10[29]) and provided multidisciplinary expert input throughout. However, it remains possible that experts with valuable complementary insights may not have been reached. Second, when reviewing the results of survey round 1, it became clear that many high scores—indicating that items should be included in CHEERS-AI—were accompanied by less-positive qualitative comments, suggesting that the items were broadly important but were not necessarily specific to AI interventions or were already captured by CHEERS 2022. This misunderstanding was rectified in survey round 2, in which participants were clearly advised that high scores should be reserved for items that were considered important to report in an EE of an AI-based intervention and may not be captured by CHEERS 2022. Furthermore, although the checklist has a dual purpose of being used by both authors and reviewers of EEs, it was piloted only in the latter context of reviewing the reporting of published evaluations, rather than guiding the reporting of a primary EE study. Because CHEERS 2022 is now often used by systematic reviewers to grade the reporting quality of identified EE studies, this can be a beneficial secondary purpose of CHEERS-AI. Finally, although some CHEERS-AI items are likely to be 'future-proofing' the reporting standards against future developments of AI in healthcare, the authors recognize that this is a rapidly evolving field. We consider that major developments should be monitored and, if needed, CHEERS-AI may need to be amended or expanded over time.

Conclusion

To date, the reporting quality of EEs of AI-enabled health technologies has been highly variable. The development of a reporting extension that supplements CHEERS 2022, to guide the reporting and reviewing of EEs of AI-enabled health interventions (CHEERS-AI), represents an important addition to the tools available to support the conduct of such evaluations. CHEERS-AI contains all pre-existing CHEERS 2022 items, including 8 elaborations to draw out important AI-specific nuances, and 10 entirely new reporting items. The use of CHEERS-AI will ensure that important AI-specific details about the intervention and their potential implications for cost-effectiveness conclusions are appropriately reported and described. We recommend that CHEERS-AI is used by researchers, editors, and reviewers conducting or assessing such EEs for dissemination or HTA purposes.

Author Disclosures

Author disclosure forms can be accessed below in the Supplemental Material section.

Supplemental Material

Supplementary data associated with this article can be found in the online version at https://doi.org/10.1016/j.jval.2024.05.006.

Article and Author Information

Accepted for Publication: May 4, 2024

Published Online: xxxx

doi: https://doi.org/10.1016/j.jval.2024.05.006

Author Affiliations: National Institute for Health and Care Excellence (NICE), England, UK (Elvidge, Hawksworth, Avşar, Chalkidou, Dawoud); Center for Health Technology Assessment and Pharmacoeconomic Research, University of Pécs, Pécs, Hungary (Zemplenyi); University of Colorado Anschutz Medical Campus, Denver, CO, USA (Zemplenyi); Syreon Research Institute, Budapest, Hungary (Zemplenyi, Petykó); Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, England, UK (Petrou); Department of Health Policy, London School of Economics and Political Science, London, England, UK (Srivastava); Biomimetics and Intelligent Systems Group, University of Oulu, Oulu, Finland (Chandra, Siirtola); EURORDIS, Brussels, Belgium (Delaye); Institute of Inflammation and Ageing, University of Birmingham, Birmingham, England, UK (Denniston); Department of Primary Care and Population Health, University College London, England, UK (Gomes); National Healthcare Institute (ZIN), Diemen, The Netherlands (Knies); Dental and Pharmaceutical Benefits Agency (TLV), Stockholm, Sweden (Nousios); Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht University, Utrecht, The Netherlands (Wang).

Correspondence: Jamie Elvidge, MSc, National Institute for Health and Care Excellence (NICE), Level 1A City Tower, Manchester, M1 4BT, United Kingdom. Email: jamie.elvidge@nice.org.uk

Author Contributions: *Concept and design:* Denniston, Zemplenyi, Hawksworth, Dawoud, Srivastava, Chandra, Elvidge, Delaye, Wang, Gomes, Siirtola, Nousios, Knies, Petrou, Petykó
Acquisition of data: Hawksworth, Elvidge
Analysis and interpretation of data: Denniston, Chalkidou, Zemplenyi, Dawoud, Srivastava, Chandra, Elvidge, Delaye, Wang, Gomes, Siirtola, Nousios, Knies, Petrou, Avşar, Petykó
Drafting of the manuscript: Denniston, Chalkidou, Srivastava, Elvidge, Avşar
Critical revision of the paper for important intellectual content: Denniston, Chalkidou, Zemplenyi, Hawksworth, Dawoud, Srivastava, Chandra, Elvidge, Delaye, Wang, Gomes, Siirtola, Nousios, Knies, Petrou, Avşar, Petykó
Provision of study materials or patients: Delaye
Obtaining funding: Dawoud
Administrative, technical, or logistic support: Elvidge, Avşar
Supervision: Dawoud
Drs Petrou, Gomes, and Dawoud are editors for *Value in Health* and had no role in the peer-review process of this article.

Funding/Support: This study has been supported by Next Generation Health Technology Assessment (HTx). The HTx project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 825162. Its main aim is to create a framework for the next generation of HTA to support patient-centered, societally oriented, and real-time decision making on access to and reimbursement for health technologies throughout Europe. This dissemination reflects only the authors' views, and the Commission is not responsible for any use that may be made of the information it contains. Stavros Petrou receives support as a UK National Institute for Health and Care Research (NIHR) Senior Investigator (NF-SI-0616-10103) and from the NIHR Applied Research Collaboration Oxford and Thames Valley.

Role of Funders/Sponsors: The European Commission had no role in the conceptualization, design, conduct or reporting of this study. This dissemination reflects only the authors' views, and the Commission is not responsible for any use that may be made of the information it contains.

Acknowledgment: The CHEERS-AI steering group would like to thank all participants in the Delphi study. The names of those who allowed acknowledgment are listed in alphabetical order: Sylvie Arbour, Carl Asche, Carolyn Ashurst, Laszlo Balkanyi, Hayley Bennett, Gerzson Boros, Rebecca Boyce, Chris Carswell, Nathorn Chaiyakunapruk, Jagpreet Chhatwal, Oriana Ciani, Gary Collins, David Dawson, David Vanness, Rossella Di Bidino, Susan Faulding, Federico Felizzi, Madeleine Haig, James Hawkins, Mikaël Hiligsmann, Annette Willemoes Holst-Kristensen, Julian Isla, Erik Koffijberg, Alexander Kostyuk, Noemi Krief, Dawn Lee, Karen Lee, Douglas Lundin, Katarzyna Markiewicz-Barreaux, Josephine Mauskopf, Karel Moons, Bertalan Németh, Guenka Petrova, Raoh-Fang (Jasmine) Pwu, Juan Carlos Rejon-Parrilla, Gabriel Rogers, Chris Sampson, Astrid Aaen Springborg, Lotte Steuten, Eric Sutherland, Jaakko Suutala, Daniel Theisen, Alexander Thompson, Lisette van Gemert-Pijnen, Thomas Walker, Ed Wilson. The authors would like to thank the EURORDIS Digital Advisory Group for reviewing the CHEERS-AI reporting items, and Veronica Popa for facilitating that review. The authors would

like to thank Harry Topps and Emma Flood for their support in developing the surveys for this study. The authors would also like to thank Don Husereau and Mike Drummond for their support during the initiation and dissemination stages of this work.

REFERENCES

1. Husereau D, Drummond M, Augustovski F, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. *BMJ*. 2022;376:e067975.
2. Medicines and Healthcare products Regulatory Agency. *Software and AI as a medical device change programme*; Published 2021. https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme. Accessed November 16, 2023.
3. National Institute for Health and Care Excellence. *Evidence standards framework (ESF) for digital health technologies*; Published 2018. https://www.nice.org.uk/corporate/ecd7. Accessed November 8, 2023.
4. National Institute for Health and Care Excellence. *Artificial intelligence-derived software to analyse chest X-rays for suspected lung cancer in primary care referrals: early value assessment*; Published 2023. https://www.nice.org.uk/guidance/hte12. Accessed November 8, 2023.
5. National Institute for Health and Care Excellence. *Artificial intelligence technologies to aid contouring for radiotherapy treatment planning: early value assessment*; Published 2023. https://www.nice.org.uk/guidance/hte11. Accessed November 8, 2023.
6. Voets MM, Veltman J, Slump CH, Siesling S, Koffijberg H. Systematic review of health economic evaluations focused on artificial intelligence in healthcare: the tortoise and the cheetah. *Value Health*. 2022;25(3):340–349.
7. Vithlani J, Hawksworth C, Elvidge J, Ayiku L, Dawoud D. Economic evaluations of artificial intelligence-based healthcare interventions: a systematic literature review of best practices in their conduct and reporting. *Front Pharmacol*. 2023;14:1220950.
8. Rivera SC, Liu X, Chan AW, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. *Nat Med*. 2020;26(9):1351–1363.
9. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. SPIRIT-AI and CONSORT-AI Working Group, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. 2020;26(9):1364–1374.
10. EQUATOR Network. *How to develop a reporting guideline*; Published June 28, 2018. https://www.equator-network.org/toolkits/developing-a-reporting-guideline/. Accessed January 30, 2023.
11. Moher D, Schulz KF, Simera I, Altman DG. Guidance for developers of Health Research reporting guidelines. *PLOS Med*. 2010;7(2):e1000217.
12. EQUATOR Network. *Reporting guidelines under development for other study designs*; Published 2023. https://www.equator-network.org/library/reporting-guidelines-under-development/reporting-guidelines-under-development-for-other-study-designs/. Accessed January 30, 2023.
13. Hawksworth C, Elvidge J, Knies S, et al. Protocol for the development of an artificial intelligence extension to the consolidated health economic evaluation reporting standards (CHEERS) 2022. *Health Econ*. Posted June 1, 2023. https://doi.org/10.1101/2023.05.31.23290788.
14. Naisola-Ruiter V. The Delphi technique: a tutorial. *Res Hosp Manag*. 2022;12(1):91–97.
15. Fitch K, Bernstein S, Aguilar M, et al. *The rand/UCLA appropriateness method User's manual*; Published 2001. https://www.rand.org/content/dam/rand/pubs/monograph_reports/2011/MR1269.pdf. Accessed February 28, 2024.
16. Hasson F, Keeney S, McKenna H. Research guidelines for the Delphi survey technique. *J Adv Nurs*. 2000;32(4):1008–1015.
17. Areia M, Mori Y, Correale L, et al. Cost-effectiveness of artificial intelligence for screening colonoscopy: a modelling study. *Lancet Digit Health*. 2022;4(6):e436–e444.
18. De Vos J, Visser LA, De Beer AA, et al. The potential cost-effectiveness of a machine learning tool that can prevent untimely Intensive Care Unit discharge. *Value Health*. 2022;25(3):359–367.
19. Hill NR, Sandler B, Mokgokong R, et al. Cost-effectiveness of targeted screening for the identification of patients with atrial fibrillation: evaluation of a machine learning risk prediction algorithm. *J Med Econ*. 2020;23(4):386–393.
20. Huang XM, Yang BF, Zheng WL, et al. Cost-effectiveness of artificial intelligence screening for diabetic retinopathy in rural China. *BMC Health Serv Res*. 2022;22(1):260.
21. Nsengiyumva NP, Hussain H, Oxlade O, et al. Triage of persons with tuberculosis symptoms using artificial intelligence–based chest radiograph interpretation: a cost-effectiveness analysis. *Open Forum Infect Dis*. 2021;8(12):ofab567.
22. Padula WV, Pronovost PJ, Makic MBF, et al. Value of hospital resources for effective pressure injury prevention: a cost-effectiveness analysis. *BMJ Qual Saf*. 2019;28(2):132–141.
23. Turino C, Benítez ID, Rafael-Palou X, et al. Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system based on machine learning aiming to improve continuous positive airway pressure treatment compliance: randomized controlled trial. *J Med Internet Res*. 2021;23(10):e24072.
24. Van Leeuwen KG, Meijer FJA, Schalekamp S, et al. Cost-effectiveness of artificial intelligence aided vessel occlusion detection in acute stroke: an early health technology assessment. *Insights Imaging*. 2021;12(1):133.
25. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health*. 2020;2(5):e240–e249.
26. Collins GS, Dhiman P, Andaur Navarro CL, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. 2021;11(7):e048008.
27. Gomez Rossi J, Rojas-Perilla N, Krois J, Schwendicke F. Cost-effectiveness of artificial intelligence as a decision-support system applied to the detection and grading of melanoma, dental caries, and diabetic retinopathy. *JAMA Netw Open*. 2022;5(3):e220269.
28. Szymanski T, Ashton R, Sekelj S, et al. Budget impact analysis of a machine learning algorithm to predict high risk of atrial fibrillation among primary care patients. *EP Eur*. 2022;24(8):1240–1247.
29. Kunst N, Siu A, Drummond M, et al. Consolidated health economic evaluation reporting standards - value of information (CHEERS-VOI): explanation and elaboration. *Value Health*. 2023;26(10):1461–1473.