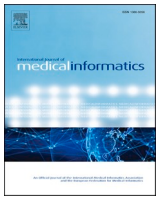
International Journal of Medical Informatics 212 (2026) 106360

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf



Review article

Healthcare practitioner involvement in data-driven clinical decision support development and evaluation: Critical narrative review of recommendations

Ruth P. Evans [a,*], Louise D. Bryant [a], Gregor Russell [a,b], David C. Wong [a], Kate Absolom [a]

[a] University of Leeds, Woodhouse Lane, Leeds LS2 9JT, UK
[b] Bradford District Care Trust, Bradford, New Mill, Victoria Road BD18 3LD, UK

ARTICLE INFO

Keywords:
Clinical decision support systems
Prediction models
Data-driven
Artificial intelligence
Machine learning
Reporting guidelines

ABSTRACT

Objective: While healthcare practitioner (HCP) involvement is widely acknowledged as essential for the development and evaluation of trustworthy data-driven clinical decision support systems (CDSS), practical guidance remains limited. This critical narrative review examines existing frameworks, highlights HCP-related considerations, and identifies areas where further guidance is warranted.

Methods: We combined searches of Ovid Medline, the EQUATOR Network Library, and relevant reviews to September 2024, seeking frameworks for developing or evaluating data-driven CDSS. Framework characteristics, coverage across the data-driven CDSS lifecycle, and details of HCP-related recommendations were extracted for analysis.

Results: 165 publications were screened, and 32 met inclusion criteria. Nine frameworks made no recommendations relating to HCP involvement. In the other 23, HCP-related recommendations were found for most phases of the data-driven CDSS development and evaluation lifecycle. Recommendations relating to HCP end users included themes of acceptability, communication, and human-AI interaction. Expert clinical input was suggested for various phases, but not required by any reporting guidelines.

Discussion: Existing guidance lacks comprehensive methods for including HCPs throughout data-driven CDSS development and evaluation. Reporting guidelines do not position HCPs as experts, which may lead to clinical expertise being overlooked. Frameworks lack detail on complex challenges such as risk communication. No frameworks suggested HCP involvement in data preparation or post-market surveillance, yet HCPs could usefully contribute to these phases.

Conclusion: HCPs should be included in data-driven CDSS development and evaluation, but there is scope to better understand how to incorporate more clinical insight, and how this might improve trustworthiness of these tools.

1. Introduction

Prediction model development is a core focus of research activity in medical and health contexts, yet few models are developed, validated and evaluated to the point where they could be harnessed for decision making in clinical settings [1].

To enable practical use, these models may be deployed as the basis of a data-driven clinical decision support system (CDSS) [2]. We use "data-driven" to encompass models derived from analysis of large patient datasets using complex statistical techniques or machine learning [2], also referred to as non-knowledge-based CDSS [3]. The CDSS inputs patient-specific data into this model to provide healthcare practitioners (HCPs) with actionable recommendations to support clinical decision making [4].

For impactful clinical implementation, collaboration with HCPs is recommended, to ensure the model relates to a clear clinical decision point [5], avoiding disconnect between "solutions" and clinical need [6]. Government bodies in the UK and USA recommend HCP involvement throughout the lifecycle of digital and data-driven health technologies, to bring multidisciplinary expertise to their development and translation to clinical settings [7,8].

Acceptance is important for HCPs adopting new tools [9], yet their

* Corresponding author.
E-mail address: R.P.Evans@leeds.ac.uk (R.P. Evans).

https://doi.org/10.1016/j.ijmedinf.2026.106360
Received 4 December 2025; Received in revised form 13 February 2026; Accepted 16 February 2026
Available online 17 February 2026

perceptions of data-driven CDSS are inconsistently sought [10], leaving key issues unresolved. For example, in machine learning or artificial intelligence (AI) predictive models, explainability is considered important for clinical user trust and translation into practice [11]. Explainable AI (XAI) is a developing field, aiming to provide insight into the underlying mechanisms behind AI model predictions [12]. However, HCPs may misplace trust in inaccurate predictions, if the explanation appears satisfactory [13] and some HCP groups place greater importance on robust evaluation demonstrating patient benefit than on explainability [14]. XAI evaluation with HCPs is inconclusive about its relationship with clinical usefulness and trust [15], and XAI studies risk clinical irrelevance by not routinely including HCP input [16].

Many reporting guidelines and frameworks exist for development and evaluation of data-driven CDSS, but there are limitations in how HCP involvement is covered. Previous reviews of reporting guidelines and frameworks [17–22] have not focused on HCP involvement, although one review of reporting guidelines and frameworks for medical AI [17] found very low coverage of stakeholder engagement.

1.1. Framework and reporting guideline definitions

Terminology and descriptions can vary [23] but for this review, a framework is a document providing structured, generalizable [17], and actionable [18] recommendations.

Reporting guidelines are a type of framework, intended to improve transparency, completeness and standardisation of published research [21,24,25]. The EQUATOR Network Library is a public collection of policy statements and guidelines for reporting health research, curated and regularly updated by experts in health research methods, statistics and reporting [26]. Such guidelines typically include a structured indication of the minimum level of detail required, via a checklist or flow diagram [24], but we also include more general guidance for reporting data-driven CDSS studies.

Although reporting guidelines are not designed as methodological frameworks, they may become a proxy where none exist, as they provide an indication of expected research processes [27].

We use “framework” as the over-arching term for all structured guidance for developing, evaluating, reporting and appraising data-driven CDSS and underlying models.

1.2. Aim and objectives

We aim to establish how published frameworks address HCP involvement in developing and evaluating data-driven CDSS. Objectives are to identify:

- The breadth of guidance available for developing and evaluating data-driven CDSS, and the purposes of existing frameworks;
- The scope and coverage of existing guidance, along the development and evaluation pipeline for data-driven CDSS;
- How existing guidance was developed, and where it was published;
- Extent and content of existing recommendations relating to HCPs, and their roles at different stages of the development and evaluation process.

2. Methods

2.1. Overview

A critical narrative review [28] was undertaken. This approach allows flexibility whilst maintaining rigour, and enables the review to go beyond description, to analysis and interpretation of included publications [29]. From the identified frameworks, recommendations about HCP involvement were extracted. These recommendations were mapped across the lifecycle phases of data-driven CDSS development and evaluation, and narratively synthesised.

2.2. Identification of frameworks

Three approaches were combined to identify frameworks relating to the development or evaluation of data-driven CDSS.

1. The EQUATOR Network Library [26] was searched to identify reporting guidelines relating to data-driven tools by selecting those tagged as suitable for “Artificial Intelligence / Machine Learning Studies” (extracted 30th August 2024).
2. Frameworks identified by a selection of recent reviews [17–22] were included as candidates for analysis.
3. A Medline search (9th September 2024), informed by the search terms used in recent reviews [17–22], was used to capture candidate frameworks published since 2021, following on from those included in previous reviews. See supplementary material for full search details.

Title and abstract screening, then full-text appraisal, was undertaken by RE, according to pre-defined criteria. Uncertainties about inclusion were resolved via discussion with a second reviewer (KA).

2.2.1. Inclusion criteria

Publications were required to be:

- Self-described as a reporting guideline or framework, or providing clear, structured recommendations for practice when developing or evaluating data-driven CDSS or underlying models.
- Broadly generalisable across clinical contexts.
- Applicable to data-driven CDSS (e.g. a prognostic model is not yet a CDSS but may become part of one in the future).
- Focused on development or evaluation (e.g. effectiveness, efficacy, acceptability).
- Available as full text and in English.

2.2.2. Exclusion criteria

- Specific recommendations for focused clinical tasks or specialist groups.
- Frameworks for other forms of digital health technology not closely related to data-driven CDSS (e.g. electronic patient records systems or applying machine learning in other areas).
- Frameworks under development.
- Frameworks focused on other topics (e.g. ethical issues or health economic evaluation).
- Reporting guidelines superseded by a more directly relevant extended version.

2.3. Data extraction and synthesis

Data on framework characteristics, their purpose, how they were created, and the phase(s) of development or evaluation covered, were extracted into a prepared form using Microsoft Excel.

Phases of development or evaluation are outlined in Table 1, based on a combination of de Hond et al [18] and Vasey et al’s [30] work, extended in response to frameworks identified in this review. Phases are not necessarily sequential, and some data-driven CDSS may omit stages or conduct some in parallel. Approval as a medical device might be assumed to fit between phases 5 and 6, or 7 and 8, but in practice may not; for instance, not all Food and Drug Administration (FDA) approved AI medical devices have been clinically validated [31] and pre-approval clinical access to innovative medical devices is proposed by the Medicines and Healthcare products Regulatory Agency (MHRA) [32].

Details of recommendations relating to HCP involvement were extracted for each framework, including sufficient text to provide context and explanation where necessary. Extracts were analysed thematically, creating a candidate list of recommendation types, which

Table 1
Phases of the data-driven CDSS lifecycle.

Phase		Description
0	Conceptualisation: defining the problem	Identifying the clinical issue that a prediction model or CDSS may address
1	Preparation of data	Including collecting, checking, cleaning, dealing with missingness
2	Model development	Deploying complex statistical or machine learning techniques on a training dataset to derive a predictive model
3	Model validation	Checking that the model remains true when tested with a new dataset
4	Software development	Creating the means for people to actually use the model to make recommendations, i.e. developing it into a CDSS
5	Data-driven CDSS evaluation – simulation	Testing the CDSS, but not in a live clinical environment
6	Data-driven CDSS evaluation – small-scale clinical	e.g. case studies, cohort studies
7	Data-driven CDSS evaluation – large-scale clinical	e.g. clinical trials
8	Data-driven CDSS implementation	Deployment to clinical use
9	Surveillance	Ongoing monitoring of a system after deployment
10	Approval as a medical device	

was then combined into themes. HCP recommendations were mapped to the phases of development and evaluation they related to, then narratively synthesised by theme and phase.

3. Results

The combined searches, after de-duplication, returned 165 records. Following title and abstract screening, 36 were assessed for eligibility, and 32 were identified for inclusion (Fig. 1).

3.1. Breadth, scope and coverage of existing frameworks

The included frameworks are summarised in Table 2.

3.1.1. Purposes of frameworks

The frameworks were categorised according to purpose. From an initial long-list, based on descriptions extracted from the publications, we identified three groups: guidance for undertaking research developing or evaluating data-driven CDSS (*methods*), *reporting* such research, and *appraising* the published outputs.

Methods (n = 15): recommendations for developers, evaluators or implementers of data-driven CDSS (or components thereof). The largest category, encompassing broad-ranging frameworks, including general guides to good practice (e.g. [33]) alongside specific methodology for elements of the process [34].

Reporting (n = 12): guidelines for researchers and developers reporting development or evaluation studies of data-driven CDSS (or components thereof). Most include a detailed checklist (n = 9), but others are less formal, without a checklist (n = 3).

Appraisal (n = 5): guidance for critical appraisal or review of studies reporting data-driven CDSS (or components thereof) development or evaluation. Some are formal checklists for researchers conducting systematic reviews [35,36], others are targeted at clinicians appraising evidence behind a new clinical tool [37,38].

Some frameworks could be included in multiple categories. For instance, Bates et al [39] cover methods and reporting, whilst Vollmer et al [40] include methods, reporting and review. Both were categorised as methods frameworks, reflecting the majority of their recommendations.

3.2. Framework development, publication, and bases for recommendations

Included frameworks were created via a wide range of methods (see Table 2). Some report detailed methods for large-scale, formal consensus building (including most formal reporting guidelines). Others are based on authors' experience, or literature reviews, case studies and expert panels or focus groups. None of the publications from government organisations include information on how those frameworks were developed [7,41–43].

Of the included frameworks, 14 (44%) were published in general medical journals, six in clinical specialty journals, and eight in medical informatics journals. The remaining four were publications from government organisations. Three of the reporting guidelines were published across multiple journals simultaneously [30,44–50].

3.3. Coverage of recommendations

Some frameworks focus on particular lifecycle phases, others span multiple phases. Table 3 provides an overview of coverage, where grey cells indicate phases addressed by a framework's overall recommendations, and coloured patterned cells indicate recommendations relating to HCPs. Where frameworks do not distinguish between evaluation phases (e.g. small- and large-scale studies [51,52]), coverage is shown in both phases.

Of the 32 frameworks, nine (28%) make no recommendations relating to HCPs [20,35,36,39,53–57], and five of these are reporting guidelines [20,54–57]. One framework does not include HCP involvement in their list of recommendations but mentions it in discussion: *"clinician involvement in model building represents an important check on variable plausibility and underlying biases…"* [39].

Analysis of HCP recommendations highlighted differences in how HCPs were characterised. Most frameworks considered HCPs in the role of end users of a prediction model or CDSS, but some characterised them as experts who could guide the work, and some positioned HCPs as both users and experts. Coverage of roles is shown by the coloured patterned cells in Table 3.

3.4. HCPs as users: Recommendations

Recommendations characterising HCPs as users are themed as: understanding who the HCP users are; human-AI interaction; acceptability of data-driven CDSS for HCP users; expertise and training; effects on HCP behaviour; communication.

3.4.1. Who the HCP users are

An understanding of who will use the data-driven CDSS could be seen as fundamental, and a simple requirement of identifying whether intended users are HCPs is noted by five frameworks: at phase 2 (model development) [58,59], phase 7 (large-scale clinical evaluation) [44,45] and phase 8 (implementation) [60].

More detail is recommended in phase 8 by the DHSC [7], which requires information about which HCPs will operate the tool.

The most detail is suggested in phase 6 (small-scale clinical evaluation) by GEP-HI [52] and DECIDE-AI [30], with the latter requiring baseline characteristics of HCP users involved in the study and recommending they represent the most common use cases and user types.

3.4.2. Human-AI interaction

With any potential automation, it should be established whether the technology or the human user controls decision making, or how they will collaborate. Although not all data-driven CDSS employs AI, seven frameworks include recommendations relating to human-AI interaction. Reporting guidelines for AI models or tools expect descriptions of how users will interact with the AI [30,44,45,58]. Responsibility for decision making is considered at conceptualisation (phase 0) [34] and detailed

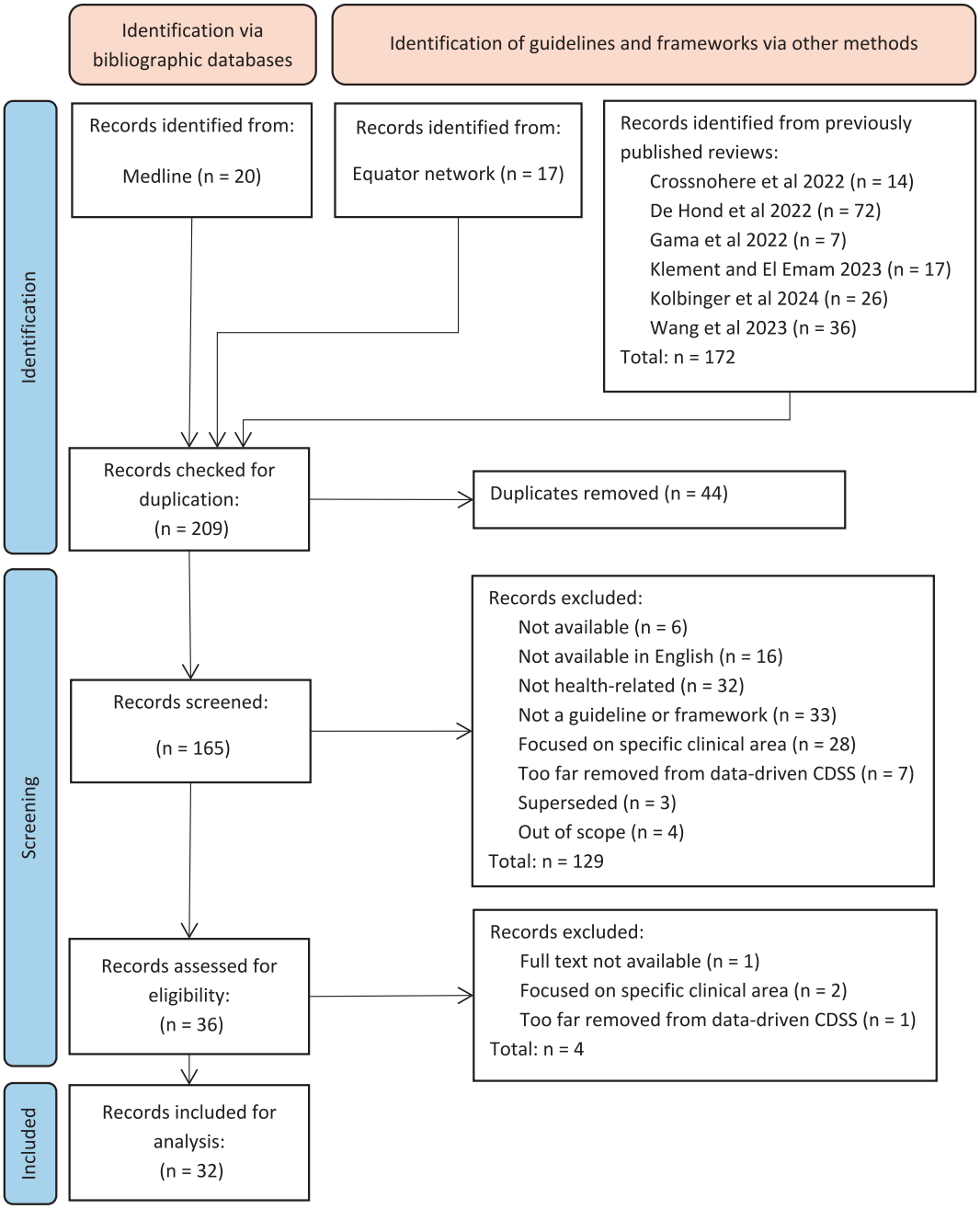


Fig. 1. PRISMA flow diagram [84].

reporting of how AI contributes to decision making is required when reporting clinical evaluation (phases 6–7) [30,44]. For approvals (phase 10), the level of professional oversight is mentioned, and performance of the human-AI team is recommended to be considered, rather than the AI alone [8].

3.4.3. Acceptability

Six frameworks make recommendations around acceptability and usability for HCPs. The DHSC requires the product to be *"…easy to use and accessible to all users."* [7] and NICE [43] highlights acceptability and credibility with UK-based HCPs, but neither provide details on what should be done to ensure or evaluate this. Other frameworks go further, with output interpretability mentioned for clinical evaluation (phases 6–7) [40], using observation of users in a simulated environment (phase 5) to *"understand why something does and doesn't work"* [61] and a recommendation to assess satisfaction over time during implementation (phase 8) [61]. DECIDE-AI (phase 6) is the only framework offering detail on usability evaluation, recommending this is carried out using established standards, and offering specific suggestions, for instance evaluation of user learning curves [30].

3.4.4. Expertise and training

Some reporting guidance expects an indication of the level of expertise required of (or training provided for) HCPs before using prediction models or data-driven CDSS. This starts at model development (phase 2) [58] and is reiterated for clinical evaluation (phases 6–7) [30,44,45] and approvals (phase 10) [43]. Two appraisal frameworks also recommend looking for information on training [38,60].

3.4.5. Effect on HCP behaviour

Five frameworks covering later phases address HCP behaviour. Two simply suggest studying interactions and effect on behaviour [41,62],

Table 2
Summary of included frameworks, ordered by framework type, then author name. Indicates the basis for the recommendations, i.e. data collection or consensus methods, the phases covered by the framework (see Table 1), and whether it includes recommendations relating to HCPs.

Author, date	Title	Framework type	Basis for recommendations (i.e. data collection or consensus building methods)	Phases covered (see table 1)	HCP recommendations included?
Bates et al, 2020 [39]	*Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence*	Methods	author opinions/experience	3, 8	None
Collin et al, 2022 [34]	*Computational Models for Clinical Applications in Personalized Medicine – Guidelines and Recommendations for Data Integration and Model Validation*	Methods	author opinions/experience	0, 1, 2, 3	Yes
Food and Drug Administration, 2017 [41]	*Software as a Medical Device (SAMD): Clinical Evaluation.*	Methods	no development methods described	6, 7, 10	Yes
Food and Drug Administration and MHRA, 2021 [42]	*Good Machine Learning Practice for Medical Device Development: Guiding Principles*	Methods	no development methods described	10	Yes
Kappen et al, 2018 [62]	*Evaluating the impact of prediction models: lessons learned, challenges, and recommendations*	Methods	case study	6, 7, 8	Yes
Moons et al, 2012 [51]	*Risk prediction models: II. External validation, model updating, and impact assessment*	Methods	author opinions/experience	3, 6, 7	Yes
Nykänen et al, 2011 [52]	*Guideline for good evaluation practice in health informatics (GEP-HI)*	Methods	expert panel and informal consensus process	6, 7	Yes
Park et al, 2020 [61]	*Evaluating artificial intelligence in medicine: phases of clinical research*	Methods	author opinions/experience	0, 2, 3, 5, 6, 7, 8	Yes
Poldrack et al, 2020 [53]	*Establishment of Best Practices for Evidence for Prediction: A Review*	Methods	author opinions/experience	2, 3	None
Smith et al, 2021 [64]	*From Code to Bedside: Implementing Artificial Intelligence Using Quality Improvement Method*	Methods	author opinions/experience	0, 8	Yes
Steyerberg and Vergouwe, 2014 [65]	*Towards better clinical prediction models: seven steps for development and an ABCD for validation*	Methods	author opinions/experience	0, 1, 2, 3	Yes
Truong et al, 2019 [33]	*A Framework for Applied AI in Healthcare*	Methods	literature review and expert panel or focus group	8	Yes
UK Department of Health and Social Care, 2021 [7]	*A guide to good practice for digital and data-driven health technologies*	Methods	no development methods described	4, 5, 6, 7, 8, 10	Yes
Vollmer et al, 2020 [40]	*Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness*	Methods	expert panel and informal consensus process	0, 1, 2, 3, 6, 7, 8, 9	Yes
Wiens et al, 2019 [63]	*Do no harm: a roadmap for responsible machine learning for health care*	Methods	author opinions/experience	2, 3, 4, 5, 6, 7, 8, 9	Yes
Collins et al, 2024 [58]	*TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods*	Reporting guidance (with checklist)	large-scale, documented, group consensus method	2, 3	Yes
Debray et al, 2023 [54]	*Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist*	Reporting guidance (with checklist)	large-scale, documented, group consensus method	2, 3	None
Klement and El Emam, 2023 [20]	*Consolidated Reporting Guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: Development and Validation*	Reporting guidance (with checklist)	literature review and expert panel or focus group	2, 3	None
Liu et al, 2020 [44]	*Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension*	Reporting guidance (with checklist)	large-scale, documented, group consensus method	7	Yes
Norgeot et al, 2020 [55]	*Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist*	Reporting guidance (with checklist)	author opinions/experience	1, 2, 3	None
Rivera et al, 2020 [45]	*Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension*	Reporting guidance (with checklist)	large-scale, documented, group consensus method	7	Yes
Snell et al, 2023 [56]	*Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA)*	Reporting guidance (with checklist)	large-scale, documented, group consensus method	3	None
Talmon et al, 2009 [57]	*STARE-HI-Statement on reporting of evaluation studies in Health Informatics*	Reporting guidance (with checklist)	large-scale, documented, group consensus method	5, 6, 7, 8, 9	None
Vasey et al, 2022 [30]	*Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI*	Reporting guidance (with checklist)	large-scale, documented, group consensus method	6	Yes
Hernandez-Boussard et al, 2020 [59]	*MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care*	Reporting guidance	author opinions/experience	1, 2, 3	Yes
National Institute for Health and Care Excellence, 2022 [43]	*Evidence standards framework for digital health technologies*	Reporting guidance	no development methods described	10	Yes

(*continued on next page*)

Table 2 (*continued*)

Author, date	Title	Framework type	Basis for recommendations (i.e. data collection or consensus building methods)	Phases covered (see table 1)	HCP recommendations included?
Stevens et al, 2020 [66]	*Recommendations for Reporting Machine Learning Analyses in Clinical Research*	Reporting guidance	author opinions/experience	1, 2, 3	Yes
Cabitza and Campagner, 2021 [60]	*The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical AI studies*	Appraisal	author opinions/experience	1, 2, 3, 8	Yes
Liu et al, 2019 [37]	*How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature*	Appraisal	author opinions/experience	3, 4, 8	Yes
Moons et al, 2014 [35]	*Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies: The CHARMS Checklist*	Appraisal	literature reviews, focus group and case studies	2, 3	None
Scott et al, 2021 [38]	*Clinician checklist for assessing suitability of machine learning applications in healthcare*	Appraisal	literature review	0, 1, 2, 3, 5, 6, 7, 8	Yes
Wolff et al, 2019 [36]	*PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies*	Appraisal	large-scale, documented, group consensus method	1, 2, 3	None

but DECIDE-AI also recommends detailing HCP agreement with the data-driven CDSS, and HCPs changing their minds based on CDSS output [30]. One framework gives study design recommendations for quantifying decision making of the HCPs [51].

One appraisal framework highlights potential effects on workflow and fatigue [37].

3.4.6. Communication

Frameworks make recommendations about four different aspects of communication with HCPs: how model outputs are presented, information about the underlying model, dissemination of evaluation results, and stakeholder awareness.

Two appraisal frameworks recommend considering how the model output (or CDSS recommendation) is presented to users, noting that using models in different ways may affect how predictions should be presented for HCPs [37] and considering whether outputs are *"clinically intelligible"* [38].

Other frameworks mention communicating to HCPs information about the underlying model, its assumptions and predictors, during clinical evaluation (phases 6–7) [62], with one framework suggesting this should be considered during model development (phase 2), to enhance HCP trust in the model [40]. The FDA recommend providing users with information on intended use, performance, training/test data and limitations [8,41]. However, they give no indication as to how this might be communicated to HCPs.

One methods framework recommends having a dissemination plan for clinical evaluation results (phase 6–7) [52].

For implementation (phase 8), one framework recommends ensuring stakeholders are aware of the value and need for the new technology [33]. Others are somewhat ambiguous as to the information being communicated, for instance *"…finding the most effective amount of information to provide to users, how and when to deliver it, and how to convey the model's confidence in its insights"* (phase 8) [61], *"Ensure that appropriate communication strategies are in place…"* (phase 10) [43] or apparently offering a warning *"…be very careful about what types of information they provide back to clinical decision makers."* (phase 8) [63].

3.5. HCPs as subject experts: Recommendations

Some frameworks position HCPs as subject experts who bring specialist clinical knowledge to the process.

Clinical input at the earliest phase is recommended by four frameworks. HCPs are *"subject matter experts"* [64] who should be consulted on the statistical goal [40]. Developers are recommended to include HCP experts early, to focus on a genuine clinical need [63], or to identify target populations and intended implementation [38].

When developing a model (phase 2), two frameworks suggest HCP experts could apply clinical domain knowledge to identify candidate variables for inclusion, or variables with limited clinical utility that may be excluded [65,66]. HCP input is also recommended for an understanding of the clinical consequences of model outputs [61].

During validation, a reference standard or "ground truth" is required to test the model against. Two appraisal frameworks identify that panels of clinical experts are likely to have determined such standards [37,38].

The software development phase is largely overlooked by the frameworks, but the DHSC highlights the benefits of HCPs and technical experts collaborating on design, testing and approval, to avoid *"…the common pitfall of having to attempt to retrofit necessary clinical input on quality and safety too late in the development process."* [7].

At the simulation stage of clinical evaluation, two frameworks suggest HCP experts could review predictions and explanations, to identify errors before testing in a live clinical environment [38,63]. This aligns with a common usability testing technique, when software designs are evaluated heuristically by a group of experts to address issues before wider usage [67].

Only one framework mentions the potential for HCP expertise in clinical evaluation, suggesting they could aid understanding of the clinical setting of an impact study trial [62]. There is only a general endorsement to include input from HCPs at the implementation stage (phase 8) [33].

For approvals, NICE recommend that developers should demonstrate that *relevant* HCPs were involved in design, development or test stages, or simply support deployment [43]. NICE require evidence that the technology is *"viewed as useful and relevant by professional experts… in the relevant field"* [43]. They also recommend that relevant HCPs validate how clinical processes are represented [43]. Similarly, the FDA require HCP expertise throughout [8].

4. Discussion

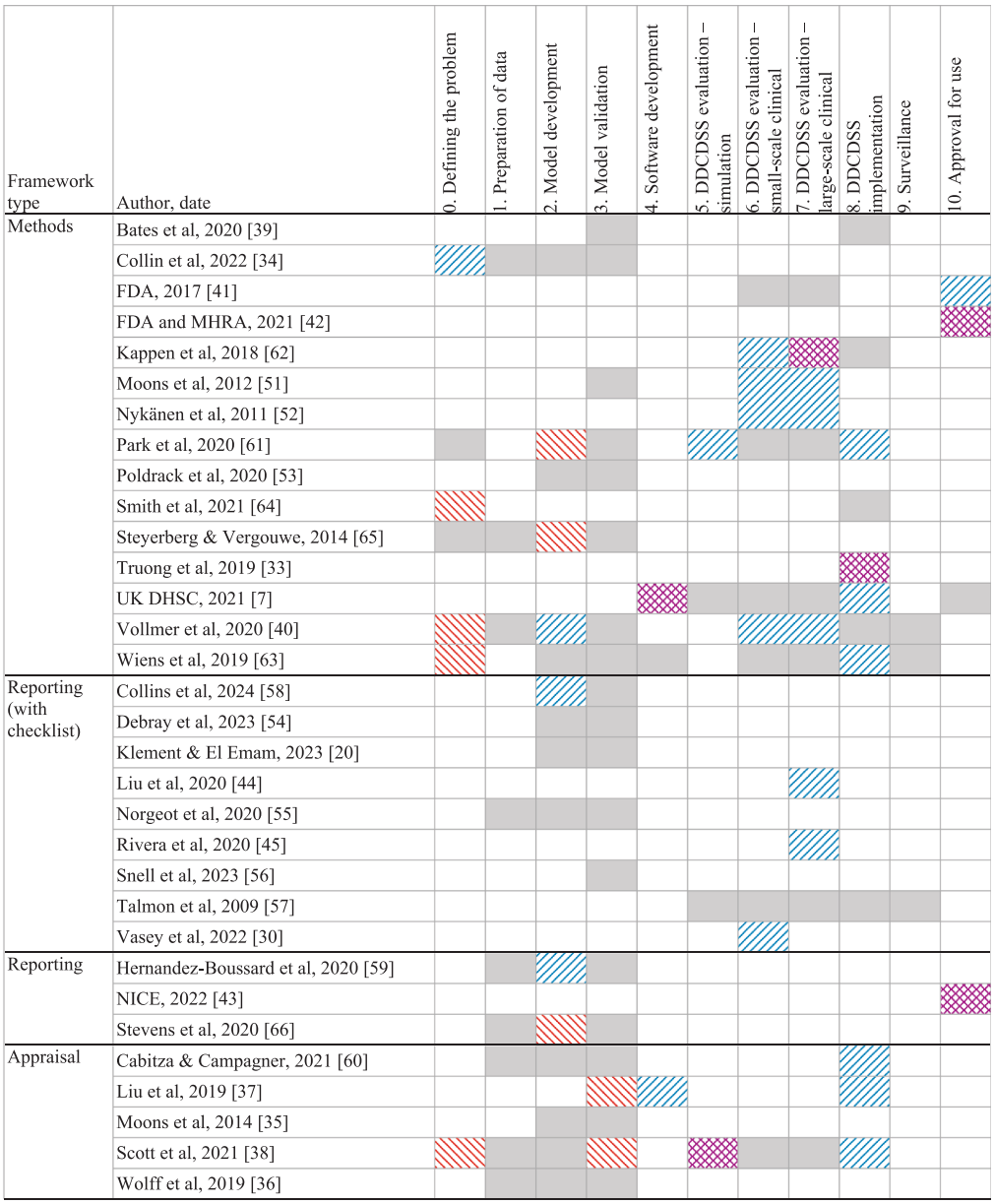
There are a variety of frameworks available for developing or evaluating data-driven CDSS and for reporting and appraising the work in research literature. Despite guidance recommending clinical input throughout [7,8], these frameworks lack specific methods for including HCPs.

4.1. Items on clinical input

A key finding was that 9/32 frameworks made no recommendations regarding HCP involvement. Formal reporting guidelines had the least coverage, with only four expecting information to be reported about HCP users [30,44,45,58], and none positioning them as experts.

Formal reporting guidelines are the most high-profile frameworks included in this review, with considerable reach and prominence due to publication in multiple high impact journals [30,44–50]. Detailed reporting guidelines may unintentionally become proxies for

Table 3
Coverage of recommendations across phases. Grey blocks indicate overall coverage of a framework; coloured blocks show HCP-specific recommendations (see key below).



methodological guidance, or misused as indicators of study quality [68]. Therefore, a lack of reporting items covering clinical input to data-driven CDSS may lead developers to overlook HCP expertise. It may also diminish the focus on comprehensive reporting of HCP involvement, leading to reduced transparency.

Frameworks considering HCP input mostly place them in the role of end users of the data-driven CDSS, but a minority frame them as clinical experts. Despite some exceptions [30,52], most frameworks do not provide or expect detail on who the HCPs are, simply expecting them to be "health professionals". "Relevant" experts are sometimes referred to (e.g. [43,64]) but this is not elaborated on.

HCPs are not a single homogenous group, so may interact with data-driven CDSS differently depending on context, specialty, and experience. Working with actual users is key for good design [69], and requirements for data-driven tools may be missed if HCPs involved in development or evaluation are not representative of the target clinical users. However, solely focusing on HCPs as end users risks missing opportunities to harness their clinical expertise throughout the

development lifecycle.

4.2. Opportunities to address gaps

The included frameworks offer no recommendations for HCP input to data preparation (phase 1), yet nine frameworks make wider recommendations here. This suggests a missed opportunity, as clinical experts with relevant domain knowledge have insights into the provenance of the data used to train and validate prediction models. This is especially important when using routinely collected health and care data, which can be highly heterogeneous across clinical settings [70], due to variation in treatment pathways, patient populations, electronic health records software, or clinical coding standards [71]. Understanding specific health care processes is essential for deriving meaningful insights from health data [72], and HCPs with experience of relevant clinical settings will be well-placed to advise on how these may impact on the proposed model. HCPs may also advise on code lists used to identify key variables in patient datasets, bringing insights into the nuances of how diagnoses and treatments might be recorded, and knowledge of disease progression and clinical pathways. HCPs may offer useful perspectives on the representativeness of datasets and applicability of resulting models to patient groups they work with; they would benefit from improved transparency of health datasets as recommended by the "STANDING Together" consensus [73].

The lack of HCP recommendations for the surveillance phase is perhaps unsurprising, as only three frameworks make wider recommendations here [40,57,63]. However, once a data-driven CDSS is approved and embedded in clinical practice, it must be monitored [74] and potentially further evaluated [75]. Clinicians could highlight concerns about digital health technologies via integrated feedback mechanisms, complementary to usual surveillance, to highlight safety concerns [76]. However, this would add to existing reporting mechanisms (e.g. [77,78]), so careful oversight would be needed to ensure HCPs are not overloaded with such expectations. There is also the further challenge of closing the loop to ensure feedback can be acted on appropriately by software providers.

4.3. Presenting information to users

We found limited coverage and a lack of detail on communication to users, both in terms of the model output (or data-driven CDSS recommendation) and the background to the model. There is recognition within frameworks that how predictions are presented to HCP users is important, but no guidance as to how best to do this, or which approaches might be preferable for different use cases.

Communication and interpretation of risk is an important and complex topic [79], and how HCPs and patients interpret recommendations from data-driven CDSS may affect shared decision making. Well-designed visual aids can improve understanding of risk and lead to better-informed decision making [80], although careful consideration is required to select the appropriate visualisation for the communication goal [81].

Although providing information about underlying models, their training data and evaluation is recommended by some frameworks, there are no suggestions for how to make this accessible and meaningful to HCPs. It is not yet known how concepts like "model facts labels" [82,83] will be presented to or received by users of data-driven CDSS.

4.4. Limitations of this review

The bibliographic database indexing of studies describing methodological, reporting and appraisal guidance is challenging to navigate, with a variety of keywords and headings used across different frameworks, so it is possible that relevant studies have been missed inadvertently by our searches. We deliberately excluded frameworks aimed at specific clinical contexts or tasks, as we sought to provide a generalisable overview.

Limiting frameworks to those published in English may have excluded relevant contributions from other languages, and we note that one systematic review [22] included frameworks from China that we were unable to access, although most were for specific clinical contexts.

This review did not seek to identify frameworks for complex interventions more generally, but these would be relevant at the implementation phase and could be included to inform future work.

5. Conclusion

There is an increasing range of guidance for researchers around how to undertake, report and appraise work developing and evaluating prediction models and data-driven CDSS. HCP-related recommendations provided by various frameworks show HCPs should be considered and involved during the process of data-driven CDSS development and evaluation.

Existing recommendations are non-contradictory and complement related work around user-centred design and implementation of complex sociotechnical interventions, yet they have notable gaps and cannot yet be amalgamated into a comprehensive practice guideline. There is considerable scope to refine and supplement recommendations to help guide meaningful contribution of HCPs and methodological approaches to support this.

With additional guidance, HCP expertise may be better harnessed throughout the process, and greater attention may be paid to specific needs of HCP users. There is a need to better understand how to incorporate relevant clinical insights into development and evaluation of data-driven CDSS, and how this might improve the trustworthiness of such tools. To create such guidance will itself require input from a wide range of HCPs and other stakeholders, and consideration of appropriate approaches to multi-disciplinary consensus-building.

AUTHOR CONTRIBUTIONS

Ruth P Evans (Conceptualisation, Methodology, Investigation, Formal analysis, Project administration, Visualisation, Writing – original draft),

Louise D Bryant (Supervision, Writing – review and editing),

Gregor Russell (Supervision, Writing – review and editing),

David C Wong (Writing – review and editing),

Kate Absolom (Conceptualisation, Methodology, Supervision, Writing – review and editing).

CRediT authorship contribution statement

Ruth P. Evans: Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Louise D. Bryant:** Writing – review & editing, Supervision. **Gregor Russell:** Writing – review & editing, Supervision. **David C. Wong:** Writing – review & editing. **Kate Absolom:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2026.106360.

References

[1] F.S. van Royen, K.G.M. Moons, G.-J. Geersing, M. van Smeden, Developing, validating, updating and judging the impact of prognostic models for respiratory diseases, Eur. Respir. J. 60 (3) (2022) 2200250, https://doi.org/10.1183/13993003.00250-2022.

[2] K. Cresswell, M. Callaghan, S. Khan, Z. Sheikh, H. Mozaffar, A. Sheikh, Investigating the use of data-driven artificial intelligence in computerised decision support systems for health and social care: a systematic review, Health Informatics J. 26 (3) (2020) 2138–2147, https://doi.org/10.1177/1460458219900452.

[3] R.T. Sutton, D. Pincock, D.C. Baumgart, D.C. Sadowski, R.N. Fedorak, K.I. Kroeker, An overview of clinical decision support systems: benefits, risks, and strategies for success, npj Digital Med. 3 (1) (2020) 17, https://doi.org/10.1038/s41746-020-0221-y.

[4] T. Koskela, S. Sandström, J. Mäkinen, H. Liira, User perspectives on an electronic decision-support tool performing comprehensive medication reviews - a focus group study with physicians and nurses, BMC Med. Inf. Decis. Making 16 (2016) 6, https://doi.org/10.1186/s12911-016-0245-z.

[5] F. Markowetz, All models are wrong and yours are useless: making clinical prediction models impactful for patients, npj Precis. Oncol. 8 (1) (2024) 54, https://doi.org/10.1038/s41698-024-00553-6.

[6] Moulds, A. and T. Horton. *Which technologies offer the biggest opportunities to save time in the NHS?* 2024 [Accessed October 2025]; Available from: https://www.health.org.uk/reports-and-analysis/briefings/which-technologies-offer-the-biggest-opportunities-to-save-time-in.

[7] Department of Health and Social Care (DHSC). *A guide to good practice for digital and data-driven health technologies*. 2021 [Accessed October 2025]; Available from: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology.

[8] US Food and Drug Administration. *Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan*. 2021 [Accessed October 2025]; Available from: https://www.fda.gov/media/145022/download.

[9] O. Perski, C.E. Short, Acceptability of digital health interventions: embracing the complexity, Transl. Behav. Med. 11 (7) (2021) 1473–1480, https://doi.org/10.1093/tbm/ibab048.

[10] R.P. Evans, L.D. Bryant, G. Russell, K. Absolom, Trust and acceptability of data-driven clinical recommendations in everyday practice: a scoping review, Int. J. Med. Inf. 183 (2024) 105342, https://doi.org/10.1016/j.ijmedinf.2024.105342.

[11] J. Amann, A. Blasimme, E. Vayena, D. Frey, V.I. Madai, and the Precise4Q consortium, Explainability for artificial intelligence in healthcare: a multidisciplinary perspective, BMC Med. Inf. Decis. Making 20 (1) (2020) 310, https://doi.org/10.1186/s12911-020-01332-6.

[12] A.F. Markus, J.A. Kors, P.R. Rijnbeek, The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies, J. Biomed. Inform. 113 (2021) 103655, https://doi.org/10.1016/j.jbi.2020.103655.

[13] B. Buijing, D. Sent, Exploring Differential Diagnosis-based Explainable AI: a Case Study in Melanoma Detection, Stud. Health Technol. Inform. 327 (2025) 507–511, https://doi.org/10.3233/shti250389.

[14] H. King, B. Williams, D. Treanor, R. Randell, How, for whom, and in what contexts will artificial intelligence be adopted in pathology? a realist interview study, J. Am. Med. Inform. Assoc. 30 (3) (2022) 529–538, https://doi.org/10.1093/jamia/ocac254.

[15] J.M. Bauer, M. Michalowski, Human-centered explainability evaluation in clinical decision-making: a critical review of the literature, J. Am. Med. Inform. Assoc. 32 (9) (2025) 1477–1484, https://doi.org/10.1093/jamia/ocaf110.

[16] Y. Abas Mohamed, B. Ee Khoo, M. Shahrimie Mohd Asaari, M. Ezane Aziz, and F. Rahiman Ghazali, Decoding the black box: Explainable AI (XAI) for cancer diagnosis, prognosis, and treatment planning-a state-of-the art systematic review, Int. J. Med. Inf. 193 (2025) 105689, https://doi.org/10.1016/j.ijmedinf.2024.105689.

[17] N.L. Crossnohere, M. Elsaid, J. Paskett, S. Bose-Brill, J.F.P. Bridges, Guidelines for Artificial Intelligence in Medicine: Literature Review and Content Analysis of Frameworks, J. Med. Internet Res. 24 (8) (2022) e36823, https://doi.org/10.2196/36823.

[18] A.A.H. de Hond, A.M. Leeuwenberg, L. Hooft, I.M.J. Kant, S.W.J. Nijman, H.J. A. van Os, J.J. Aardoom, T.P.A. Debray, E. Schuit, M. van Smeden, J.B. Reitsma, E. W. Steyerberg, N.H. Chavannes, K.G.M. Moons, Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review, npj Digital Med. 5 (1) (2022) 2, https://doi.org/10.1038/s41746-021-00549-7.

[19] F. Gama, D. Tyskbo, J. Nygren, J. Barlow, J. Reed, P. Svedberg, Implementation Frameworks for Artificial Intelligence translation into Health Care Practice: Scoping Review, J. Med. Internet Res. 24 (1) (2022) e32215, https://doi.org/10.2196/32215.

[20] W. Klement, K. El Emam, Consolidated Reporting guidelines for Prognostic and Diagnostic Machine Learning Modeling Studies: Development and Validation, J. Med. Internet Res. 25 (2023) e48763, https://doi.org/10.2196/48763.

[21] F.R. Kolbinger, G.P. Veldhuizen, J. Zhu, D. Truhn, J.N. Kather, Reporting guidelines in medical artificial intelligence: a systematic review and meta-analysis, Communications Medicine 4 (1) (2024) 71, https://doi.org/10.1038/s43856-024-00492-0.

[22] Y. Wang, N. Li, L. Chen, M. Wu, S. Meng, Z. Dai, Y. Zhang, M. Clarke, Guidelines, Consensus statements, and Standards for the use of Artificial Intelligence in Medicine: Systematic Review, J. Med. Internet Res. 25 (2023) e46089, https://doi.org/10.2196/46089.

[23] P. Nilsen, Making sense of implementation theories, models and frameworks, Implement. Sci. 10 (1) (2015) 53, https://doi.org/10.1186/s13012-015-0242-0.

[24] M.M. Schlussel, M.K. Sharp, J.A. de Beyer, S. Kirtley, P. Logullo, P. Dhiman, A. MacCarthy, A. Koroleva, B. Speich, G.S. Bullock, D. Moher, G.S. Collins, Reporting guidelines used varying methodology to develop recommendations, J. Clin. Epidemiol. 159 (2023) 246–256, https://doi.org/10.1016/j.jclinepi.2023.03.018.

[25] Equator Network. *What is a reporting guideline?* 2025 [Accessed 18th February 2025]; Available from: https://www.equator-network.org/about-us/what-is-a-reporting-guideline/.

[26] Equator Network. *Equator Network Reporting Guidelines*. 2025 [Accessed 18th February 2025]; Available from: https://www.equator-network.org/reporting-guidelines/.

[27] D.G. Altman, I. Simera, *Using Reporting guidelines Effectively to Ensure Good Reporting of Health Research*, in *guidelines for Reporting Health Research*, A User's Manual. (2014) 32–40.

[28] A. Sutton, M. Clowes, L. Preston, A. Booth, Meeting the review family: exploring review types and associated information retrieval requirements, Health Information & Libraries Journal 36 (3) (2019) 202–222, https://doi.org/10.1111/hir.12276.

[29] M.J. Grant, A. Booth, A typology of reviews: an analysis of 14 review types and associated methodologies, Health Information & Libraries Journal 26 (2) (2009) 91–108, https://doi.org/10.1111/j.1471-1842.2009.00848.x.

[30] B. Vasey, M. Nagendran, B. Campbell, D.A. Clifton, G.S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, X. Liu, B.A. Mateen, P. Mathur, M. D. McCradden, L. Morgan, J. Ordish, C. Rogers, S. Saria, D.S.W. Ting, P. Watkinson, W. Weber, P. Wheatstone, P. McCulloch, Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, BMJ 377 (2022) e070904, https://doi.org/10.1136/bmj-2022-070904.

[31] S. Chouffani El Fassi, A. Abdullah, Y. Fang, S. Natarajan, A.B. Masroor, N. Kayali, S. Prakash, G.E. Henderson, Not all AI health tools with regulatory authorization are clinically validated, Nat. Med. 30 (10) (2024) 2718–2720, https://doi.org/10.1038/s41591-024-03203-3.

[32] Medicines and Healthcare Products Regulatory Agency. *Statement of Policy Intent: Early Access to Innovative Medical Devices*. 2025 [Accessed September 2025]; Available from: https://www.gov.uk/government/publications/statement-of-policy-intent-early-access-to-innovative-medical-devices/statement-of-policy-intent-early-access-to-innovative-medical-devices.

[33] T. Truong, P. Gilbank, K. Johnson-Cover, A. Ieraci, A Framework for Applied AI in Healthcare, Stud. Health Technol. Inform. 264 (2019) 1993–1994, https://doi.org/10.3233/shti190751.

[34] C.B. Collin, T. Gebhardt, M. Golebiewski, T. Karaderi, M. Hillemanns, F.M. Khan, A. Salehzadeh-Yazdi, M. Kirschner, S. Krobitsch, E.-S.P. consortium, and L. Kuepfer, Computational Models for Clinical applications in Personalized Medicine—Guidelines and Recommendations for Data Integration and Model Validation. *Journal of*, Pers. Med. 12 (2) (2022) 166, https://doi.org/10.3390/jpm12020166.

[35] K.G.M. Moons, J.A.H. de Groot, W. Bouwmeester, Y. Vergouwe, S. Mallett, D. G. Altman, J.B. Reitsma, G.S. Collins, Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies, The CHARMS Checklist. PLOS Medicine 11 (10) (2014) e1001744, https://doi.org/10.1371/journal.pmed.1001744.

[36] R.F. Wolff, K.G.M. Moons, R.D. Riley, P.F. Whiting, M. Westwood, G.S. Collins, J. B. Reitsma, J. Kleijnen, S. Mallett, PROBAST: a Tool to Assess the risk of Bias and Applicability of Prediction Model Studies, Ann. Intern. Med. 170 (1) (2019) 51–58, https://doi.org/10.7326/M18-1376.

[37] Y. Liu, P.-H.-C. Chen, J. Krause, L. Peng, How to Read Articles that use Machine Learning: users' Guides to the Medical Literature, JAMA 322 (18) (2019) 1806–1816, https://doi.org/10.1001/jama.2019.16489.

[38] I. Scott, S. Carter, E. Coiera, Clinician checklist for assessing suitability of machine learning applications in healthcare, BMJ Health & Care Informatics 28 (1) (2021) e100251, https://doi.org/10.1136/bmjhci-2020-100251.

[39] D.W. Bates, A. Auerbach, P. Schulam, A. Wright, S. Saria, Reporting and Implementing Interventions Involving Machine Learning and Artificial Intelligence, Ann. Intern. Med. 172 (11_Supplement) (2020) S137–S144, https://doi.org/10.7326/m19-0872.

[40] S. Vollmer, B.A. Mateen, G. Bohner, F.J. Király, R. Ghani, P. Jonsson, S. Cumbers, A. Jonas, K.S.L. McAllister, P. Myles, D. Grainger, M. Birse, R. Branson, K.G. M. Moons, G.S. Collins, J.P.A. Ioannidis, C. Holmes, H. Hemingway, Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness, BMJ 368 (2020) l6927, https://doi.org/10.1136/bmj.l6927.

[41] US Food and Drug Administration. *Software as a Medical Device (SAMD): Clinical Evaluation*. 2017 [Accessed October 2025]; Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/software-medical-device-samd-clinical-evaluation.

[42] US Food and Drug Administration, Health Canada, and Medicines and Healthcare Products Regulatory Agency. *Good Machine Learning Practice for Medical Device Development: Guiding Principles*. 2021 [Accessed October 2025]; Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

[43] National Institute for Health and Care Excellence. *Evidence standards framework for digital health technologies*. 2022 [Accessed October 2025]; Available from: https://www.nice.org.uk/corporate/ecd7.

[44] X. Liu, S.C. Rivera, D. Moher, M.J. Calvert, A.K. Denniston, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension, BMJ 370 (2020) m3164, https://doi.org/10.1136/bmj.m3164.

[45] S.C. Rivera, X. Liu, A.-W. Chan, A.K. Denniston, M.J. Calvert, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension, BMJ 370 (2020) m3210, https://doi.org/10.1136/bmj.m3210.

[46] B. Vasey, M. Nagendran, B. Campbell, D.A. Clifton, G.S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, X. Liu, B.A. Mateen, P. Mathur, M. D. McCradden, L. Morgan, J. Ordish, C. Rogers, S. Saria, D.S.W. Ting, P. Watkinson, W. Weber, P. Wheatstone, P. McCulloch, A.Y. Lee, A.G. Fraser, A. Connell, A. Vira, A. Esteva, A.D. Althouse, A.L. Beam, A. de Hond, A.-L. Boulesteix, A. Bradlow, A. Ercole, A. Paez, A. Tsanas, B. Kirby, B. Glocker, C. Velardo, C.M. Park, C. Hehakaya, C. Baber, C. Paton, C. Johner, C.J. Kelly, C.J. Vincent, C. Yau, C. McGenity, C. Gatsonis, C. Faivre-Finn, C. Simon, D. Sent, D. Bzdok, D. Treanor, D.C. Wong, D.F. Steiner, D. Higgins, D. Benson, D.P. O'Regan, D.V. Gunasekaran, D. Danks, E. Neri, E. Kyrimi, F. Schwendicke, F. Magrabi, F. Ives, F.E. Rademakers, G.E. Fowler, G. Frau, H.D.J. Hogg, H.J. Marcus, H.-P. Chan, H. Xiang, H. F. McIntyre, H. Harvey, H. Kim, I. Habli, J.C. Fackler, J. Shaw, J. Higham, J. M. Wohlgemut, J. Chong, J.-E. Bibault, J.F. Cohen, J. Kers, J. Morley, J. Krois, J. Monteiro, J. Horovitz, J. Fletcher, J. Taylor, J.H. Yoon, K. Singh, K.G.M. Moons, K. Karpathakis, K. Catchpole, K. Hood, K. Balaskas, K. Kamnitsas, L. Militello, L. Wynants, L. Oakden-Rayner, L.B. Lovat, L.J.M. Smits, L.C. Hinske, M. K. ElZarrad, M. van Smeden, M. Giavina-Bianchi, M. Daley, M.P. Sendak, M. Sujan, M. Rovers, M. DeCamp, M. Woodward, M. Komorowski, M. Marsden, M. Mackintosh, M.D. Abramoff, M.Á.A. de la Hoz, N. Hambidge, N. Daly, N. Peek, O. Redfern, O.F. Ahmad, P.M. Bossuyt, P.A. Keane, P.N.P. Ferreira, P. Schnell-Inderst, P. Mascagni, P. Dasgupta, P. Guan, R. Barnett, R. Kader, R. Chopra, R. M. Mann, R. Sarkar, S.M. Mäenpää, S.G. Finlayson, S. Vollam, S.J. Vollmer, S. H. Park, S. Laher, S. Joshi, S.L. van der Meijden, S.C. Shelmerdine, T.-E. Tan, T.J. W. Stocker, V. Giannini, V.I. Madai, V. Newcombe, W.Y. Ng, W.A. Rogers, W. Ogallo, Y. Park, Z.b., Perkins and the Decide-AI expert group, Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI, Nat. Med. 28 (5) (2022) 924–933, https://doi.org/10.1038/s41591-022-01772-9.

[47] Liu, X., S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, A.-W. Chan, A. Darzi, C. Holmes, C. Yau, H. Ashrafian, J.J. Deeks, L. Ferrante di Ruffano, L. Faes, P.A. Keane, S.J. Vollmer, A.Y. Lee, A. Jonas, A. Esteva, A.L. Beam, A.-W. Chan, M. B. Panico, C.S. Lee, C. Haug, C.J. Kelly, C. Yau, C. Mulrow, C. Espinoza, J. Fletcher, D. Paltoo, E. Manna, G. Price, G.S. Collins, H. Harvey, J. Matcham, J. Monteiro, M. K. ElZarrad, L. Ferrante di Ruffano, L. Oakden-Rayner, M. McCradden, P.A. Keane, R. Savage, R. Golub, R. Sarkar, S. Rowley, S.-A. The, C.-A.W. Group, A.I. Spirit, C.-A.S. Group, A.I. Spirit, and C.-A.C. Group, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension, Nat. Med. 26 (9) (2020) 1364–1374, https://doi.org/10.1038/s41591-020-1034-x.

[48] X. Liu, S. Cruz Rivera, D. Moher, M.J. Calvert, A.K. Denniston, H. Ashrafian, A. L. Beam, A.-W. Chan, G.S. Collins, A.D.J. Deeks, M.K. ElZarrad, C. Espinoza, A. Esteva, L. Faes, L. Ferrante di Ruffano, J. Fletcher, R. Golub, H. Harvey, C. Haug, C. Holmes, A. Jonas, P.A. Keane, C.J. Kelly, A.Y. Lee, C.S. Lee, E. Manna, J. Matcham, M. McCradden, J. Monteiro, C. Mulrow, L. Oakden-Rayner, D. Paltoo, M.B. Panico, G. Price, S. Rowley, R. Savage, R. Sarkar, S.J. Vollmer, C. Yau, Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *the Lancet Digital*, Health 2 (10) (2020) e537–e548, https://doi.org/10.1016/S2589-7500(20)30218-1.

[49] S. Cruz Rivera, X. Liu, A.-W. Chan, A.K. Denniston, M.J. Calvert, H. Ashrafian, A. L. Beam, G.S. Collins, A. Darzi, J.J. Deeks, M.K. ElZarrad, C. Espinoza, A. Esteva, L. Faes, L. Ferrante di Ruffano, J. Fletcher, R. Golub, H. Harvey, C. Haug, C. Holmes, A. Jonas, P.A. Keane, C.J. Kelly, A.Y. Lee, C.S. Lee, E. Manna, J. Matcham, M. McCradden, D. Moher, J. Monteiro, C. Mulrow, L. Oakden-Rayner, D. Paltoo, M.B. Panico, G. Price, S. Rowley, R. Savage, R. Sarkar, S.J. Vollmer, C. Yau, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *the Lancet Digital*, Health 2 (10) (2020) e549–e560, https://doi.org/10.1016/S2589-7500(20)30219-3.

[50] Cruz Rivera, S., X. Liu, A.-W. Chan, A.K. Denniston, M.J. Calvert, A. Darzi, C. Holmes, C. Yau, D. Moher, H. Ashrafian, J.J. Deeks, L. Ferrante di Ruffano, L. Faes, P.A. Keane, S.J. Vollmer, A.Y. Lee, A. Jonas, A. Esteva, A.L. Beam, M.B. Panico, C.S. Lee, C. Haug, C.J. Kelly, C. Yau, C. Mulrow, C. Espinoza, J. Fletcher, D. Moher, D. Paltoo, E. Manna, G. Price, G.S. Collins, H. Harvey, J. Matcham, J. Monteiro, M.K. ElZarrad, L. Ferrante di Ruffano, L. Oakden-Rayner, M. McCradden, P.A. Keane, R. Savage, R. Golub, R. Sarkar, S. Rowley, S.-A. The, C.-A.W. Group, A.I. Spirit, C.-A.S. Group, A.I. Spirit, and C.-A.C. Group, Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension, Nat. Med. 26 (9) (2020) 1351–1363, https://doi.org/10.1038/s41591-020-1037-7.

[51] K.G.M. Moons, A.P. Kengne, D.E. Grobbee, P. Royston, Y. Vergouwe, D.G. Altman, M. Woodward, Risk prediction models: II. External validation, model updating, and impact assessment, Heart 98 (9) (2012) 691–698, https://doi.org/10.1136/heartjnl-2011-301247.

[52] P. Nykänen, J. Brender, J. Talmon, N. de Keizer, M. Rigby, M.-C. Beuscart-Zephir, E. Ammenwerth, Guideline for good evaluation practice in health informatics (GEP-HI), Int. J. Med. Inf. 80 (12) (2011) 815–827, https://doi.org/10.1016/j.ijmedinf.2011.08.004.

[53] R.A. Poldrack, G. Huckins, G. Varoquaux, Establishment of best Practices for evidence for Prediction: a Review, JAMA Psychiat. 77 (5) (2020) 534–540, https://doi.org/10.1001/jamapsychiatry.2019.3671.

[54] T.P.A. Debray, G.S. Collins, R.D. Riley, K.I.E. Snell, B. Van Calster, J.B. Reitsma, K. G.M. Moons, Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist, BMJ 380 (2023) e071018, https://doi.org/10.1136/bmj-2022-071018.

[55] B. Norgeot, G. Quer, B.K. Beaulieu-Jones, A. Torkamani, R. Dias, M. Gianfrancesco, R. Arnaout, I.S. Kohane, S. Saria, E. Topol, Z. Obermeyer, B. Yu, A.J. Butte, Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist, Nat. Med. 26 (9) (2020) 1320–1324, https://doi.org/10.1038/s41591-020-1041-y.

[56] K.I.E. Snell, B. Levis, J.A.A. Damen, P. Dhiman, T.P.A. Debray, L. Hooft, J. B. Reitsma, K.G.M. Moons, G.S. Collins, R.D. Riley, Transparent reporting of multivariable prediction models for individual prognosis or diagnosis: checklist for systematic reviews and meta-analyses (TRIPOD-SRMA), BMJ 381 (2023) e073538, https://doi.org/10.1136/bmj-2022-073538.

[57] J. Talmon, E. Ammenwerth, J. Brender, N. de Keizer, P. Nykänen, M. Rigby, STARE-HI—Statement on reporting of evaluation studies in Health Informatics, Int. J. Med. Inf. 78 (1) (2009) 1–9, https://doi.org/10.1016/j.ijmedinf.2008.09.002.

[58] G.S. Collins, K.G.M. Moons, P. Dhiman, R.D. Riley, A.L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J.B. Reitsma, M. van Smeden, A.-L. Boulesteix, J. C. Camaradou, L.A. Celi, S. Denaxas, A.K. Denniston, B. Glocker, R.M. Golub, H. Harvey, G. Heinze, M.M. Hoffman, A.P. Kengne, E. Lam, N. Lee, E.W. Loder, L. Maier-Hein, B.A. Mateen, M.D. McCradden, L. Oakden-Rayner, J. Ordish, R. Parnell, S. Rose, K. Singh, L. Wynants, P. Logullo, TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378, https://doi.org/10.1136/bmj-2023-078378.

[59] T. Hernandez-Boussard, S. Bozkurt, J.P.A. Ioannidis, N.H. Shah, MINIMAR (MINimum Information for Medical AI Reporting): developing reporting standards for artificial intelligence in health care, J. Am. Med. Inform. Assoc. 27 (12) (2020) 2011–2015, https://doi.org/10.1093/jamia/ocaa088.

[60] F. Cabitza, A. Campagner, The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies, Int. J. Med. Inf. 153 (2021) 104510, https://doi.org/10.1016/j.ijmedinf.2021.104510.

[61] Y. Park, G.P. Jackson, M.A. Foreman, D. Gruen, J. Hu, A.K. Das, Evaluating artificial intelligence in medicine: phases of clinical research, JAMIA Open 3 (3) (2020) 326–331, https://doi.org/10.1093/jamiaopen/ooaa033.

[62] T.H. Kappen, W.A. van Klei, L. van Wolfswinkel, C.J. Kalkman, Y. Vergouwe, K.G. M. Moons, Evaluating the impact of prediction models: lessons learned, challenges, and recommendations, Diagn. Progn. Res. 2 (1) (2018) 11, https://doi.org/10.1186/s41512-018-0033-6.

[63] J. Wiens, S. Saria, M. Sendak, M. Ghassemi, V.X. Liu, F. Doshi-Velez, K. Jung, K. Heller, D. Kale, M. Saeed, P.N. Ossorio, S. Thadaney-Israni, A. Goldenberg, Do no harm: a roadmap for responsible machine learning for health care, Nat. Med. 25 (9) (2019) 1337–1340, https://doi.org/10.1038/s41591-019-0548-6.

[64] M. Smith, A. Sattler, G. Hong, S. Lin, From Code to Bedside: Implementing Artificial Intelligence using Quality Improvement Methods, J. Gen. Intern. Med. 36 (4) (2021) 1061–1066, https://doi.org/10.1007/s11606-020-06394-w.

[65] E.W. Steyerberg, Y. Vergouwe, Towards better clinical prediction models: seven steps for development and an ABCD for validation, Eur. Heart J. 35 (29) (2014) 1925–1931, https://doi.org/10.1093/eurheartj/ehu207.

[66] L.M. Stevens, B.J. Mortazavi, R.C. Deo, L. Curtis, D.P. Kao, Recommendations for Reporting Machine Learning analyses in Clinical Research, Circ. Cardiovasc. Qual. Outcomes 13 (10) (2020) e006556, https://doi.org/10.1161/circoutcomes.120.006556.

[67] M.C. Neff, D. Schütze, S. Holtz, S.M. Köhler, J. Vasseur, N. Ahmadi, H. Storf, J. Schaaf, Development and expert inspections of the user interface for a primary care decision support system, Int. J. Med. Inf. 192 (2024) 105651, https://doi.org/10.1016/j.ijmedinf.2024.105651.

[68] K.F. Schulz, D. Moher, D.G. Altman, *Ambiguities and Confusions between Reporting and Conduct*, in *guidelines for Reporting Health Research*, A User's Manual. (2014) 41–47.

[69] D.A. Norman, *The Design of Everyday Things*, Basic Books, USA, 2002.

[70] D. Kotecha, F.W. Asselbergs, S. Achenbach, S.D. Anker, D. Atar, C. Baigent, A. Banerjee, B. Beger, G. Brobert, B. Casadei, C. Ceccarelli, M.R. Cowie, F. Crea, M. Cronin, S. Denaxas, A. Derix, D. Fitzsimons, M. Fredriksson, C.P. Gale, G. V. Gkoutos, W. Goettsch, H. Hemingway, M. Ingvar, A. Jonas, R. Kazmierski, S. Løgstrup, R.T. Lumbers, T.F. Lüscher, P. McGreavy, I.L. Piña, L. Roessig, C. Steinbeisser, M. Sundgren, B. Tyl, G. van Thiel, K. van Bochove, P.E. Vardas, T. Villanueva, M. Vrana, W. Weber, F. Weidinger, S. Windecker, A. Wood, D. E. Grobbee, CODE-EHR best practice framework for the use of structured electronic healthcare records in clinical research, BMJ 378 (2022) e069048, https://doi.org/10.1136/bmj-2021-069048.

[71] A.S. Tang, S.R. Woldemariam, S. Miramontes, B. Norgeot, T.T. Oskotsky, M. Sirota, Harnessing EHR data for health research, Nat. Med. 30 (7) (2024) 1847–1855, https://doi.org/10.1038/s41591-024-03074-8.

[72] D. Agniel, I.S. Kohane, G.M. Weber, Biases in electronic health record data due to processes within the healthcare system: retrospective observational study, BMJ 361 (2018) k1479, https://doi.org/10.1136/bmj.k1479.

[73] J.E. Alderman, J. Palmer, E. Laws, M.D. McCradden, J. Ordish, M. Ghassemi, S. R. Pfohl, N. Rostamzadeh, H. Cole-Lewis, B. Glocker, M. Calvert, T.J. Pollard, J. Gill, J. Gath, A. Adebajo, J. Beng, C.H. Leung, S. Kuku, L.-A. Farmer, R.N. Matin, B.A. Mateen, F. McKay, K. Heller, A. Karthikesalingam, D. Treanor, M. Mackintosh, L. Oakden-Rayner, R. Pearson, A.K. Manrai, P. Myles, J. Kumuthini, Z. Kapacee, N. J. Sebire, L.H. Nazer, J. Seah, A. Akbari, L. Berman, J.W. Gichoya, L. Righetto, D. Samuel, W. Wasswa, M. Charalambides, A. Arora, S. Pujari, C. Summers, E. Sapey, S. Wilkinson, V. Thakker, A. Denniston, X. Liu, Tackling algorithmic bias and promoting transparency in health datasets: the STANDING together consensus recommendations, The Lancet Digital Health 7 (1) (2025) e64–e88, https://doi.org/10.1016/S2589-7500(24)00224-3.

[74] European Union, *Regulation (EU) 2017/645 of the European Parliament and of the Council of 5 April 2017 on medical devices (Article 83)*, European Union, Editor. 2017.

[75] A. Abbasi, D. Rivera, L.H. Curtis, R.M. Califf, Post-approval evidence generation: a shared responsibility for healthcare, Nat. Med. 30 (11) (2024) 3046–3049, https://doi.org/10.1038/s41591-024-03241-x.

[76] R. Mathias, B. Vasey, A. Chalkidou, L. Riedemann, T. Melvin, S. Gilbert, Safe AI-enabled digital health technologies need built-in open feedback, Nat. Med. 31 (2) (2025) 370–375, https://doi.org/10.1038/s41591-024-03397-6.

[77] NHS England. *Learn from patient safety events (LFPSE) service*. 2025 [Accessed 25th July 2025]; Available from: https://www.england.nhs.uk/patient-safety/patient-safety-insight/learning-from-patient-safety-events/learn-from-patient-safety-events-service/.

[78] US Food and Drug Administration. *Manufacturer and User Facility Device Experience (MAUDE) Database*. 2025 [Accessed 25th July 2025]; Available from: https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfmaude/search.cfm.

[79] J.S. Ancker, N.C. Benda, M.M. Sharma, S.B. Johnson, M. Demetres, D. Delgado, B. J. Zikmund-Fisher, Scope, Methods, and Overview Findings for the making Numbers Meaningful evidence Review of Communicating Probabilities in Health: a Systematic Review, MDM Policy & Practice 10 (1) (2025) 23814683241255334, https://doi.org/10.1177/23814683241255334.

[80] R. Garcia-Retamero, E.T. Cokely, Designing Visual Aids that Promote Risk Literacy: a Systematic Review of Health Research and Evidence-based Design Heuristics, Hum. Factors 59 (4) (2017) 582–627, https://doi.org/10.1177/0018720817690634.

[81] J.S. Ancker, N.C. Benda, B.J. Zikmund-Fisher, Do you want to promote recall, perceptions, or behavior? the best data visualization depends on the communication goal, J. Am. Med. Inform. Assoc. 31 (2) (2023) 525–530, https://doi.org/10.1093/jamia/ocad137.

[82] M.P. Sendak, M. Gao, N. Brajer, S. Balu, Presenting machine learning model information to clinical end users with model facts labels, npj Digital Med. 3 (1) (2020) 41, https://doi.org/10.1038/s41746-020-0253-3.

[83] K.A. Bramstedt, Artificial Intelligence (AI) Facts Labels: an innovative Disclosure Tool Promoting Patient-Centric Transparency in Healthcare AI Systems, J. Med. Syst. 49 (1) (2025) 78, https://doi.org/10.1007/s10916-025-02216-w.

[84] M.J. Page, J.E. McKenzie, P.M. Bossuyt, I. Boutron, T.C. Hoffmann, C.D. Mulrow, L. Shamseer, J.M. Tetzlaff, E.A. Akl, S.E. Brennan, R. Chou, J. Glanville, J. M. Grimshaw, A. Hróbjartsson, M.M. Lalu, T. Li, E.W. Loder, E. Mayo-Wilson, S. McDonald, L.A. McGuinness, L.A. Stewart, J. Thomas, A.C. Tricco, V.A. Welch, P. Whiting, D. Moher, The PRISMA 2020 statement: an updated guideline for reporting systematic reviews, Syst. Rev. 10 (1) (2021) 89, https://doi.org/10.1186/s13643-021-01626-4.