Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment? A systematic review

Line Farah, Julie Davaze-Schneider, Tess Martin, Pierre Nguyen, Isabelle Borget, Nicolas Martelli

▶ **To cite this version:**

Line Farah, Julie Davaze-Schneider, Tess Martin, Pierre Nguyen, Isabelle Borget, et al.. Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment? A systematic review. Artificial Intelligence in Medicine, 2023, 140, 10.1016/j.artmed.2023.102547 . hal-04106309

HAL Id: hal-04106309
https://hal.science/hal-04106309v1

Submitted on 2 Jun 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NoDerivatives 4.0 International License

Artificial Intelligence In Medicine 140 (2023) 102547

Contents lists available at ScienceDirect

Artificial Intelligence In Medicine

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

Are current clinical studies on artificial intelligence-based medical devices comprehensive enough to support a full health technology assessment? A systematic review

Line Farah [a,b,*], Julie Davaze-Schneider [c], Tess Martin [a,c], Pierre Nguyen [c], Isabelle Borget [a,d,e], Nicolas Martelli [a,c]

[a] Groupe de Recherche et d'accueil en Droit et Economie de la Santé (GRADES) Department, University Paris-Saclay, Orsay, France
[b] Innovation Center for Medical Devices, Foch Hospital, 40 Rue Worth, 92150 Suresnes, France
[c] Pharmacy Department, Georges Pompidou European Hospital, AP-HP, 20 Rue Leblanc, 75015 Paris, France
[d] Department of Biostatistics and Epidemiology, Gustave Roussy, University Paris-Saclay, 94805 Villejuif, France
[e] Oncostat U1018, Inserm, University Paris-Saclay, Équipe Labellisée Ligue Contre le Cancer, Villejuif, France

ARTICLE INFO

Keywords:
Artificial intelligence
Machine learning
Artificial intelligence-based medical device
Health technology assessment
Clinical trial
Economic evaluation

ABSTRACT

Introduction: Artificial Intelligence-based Medical Devices (AI-based MDs) are experiencing exponential growth in healthcare. This study aimed to investigate whether current studies assessing AI contain the information required for health technology assessment (HTA) by HTA bodies.

Methods: We conducted a systematic literature review based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses methodology to extract articles published between 2016 and 2021 related to the assessment of AI-based MDs. Data extraction focused on study characteristics, technology, algorithms, comparators, and results. AI quality assessment and HTA scores were calculated to evaluate whether the items present in the included studies were concordant with the HTA requirements. We performed a linear regression for the HTA and AI scores with the explanatory variables of the impact factor, publication date, and medical specialty. We conducted a univariate analysis of the HTA score and a multivariate analysis of the AI score with an alpha risk of 5 %.

Results: Of 5578 retrieved records, 56 were included. The mean AI quality assessment score was 67 %; 32 % of articles had an AI quality score $\geq$ 70 %, 50 % had a score between 50 % and 70 %, and 18 % had a score under 50 %. The highest quality scores were observed for the study design (82 %) and optimisation (69 %) categories, whereas the scores were lowest in the clinical practice category (23 %). The mean HTA score was 52 % for all seven domains. 100 % of the studies assessed clinical effectiveness, whereas only 9 % evaluated safety, and 20 % evaluated economic issues. There was a statistically significant relationship between the impact factor and the HTA and AI scores (both $p = 0.046$).

Discussion: Clinical studies on AI-based MDs have limitations and often lack adapted, robust, and complete evidence. High-quality datasets are also required because the output data can only be trusted if the inputs are reliable. The existing assessment frameworks are not specifically designed to assess AI-based MDs. From the perspective of regulatory authorities, we suggest that these frameworks should be adapted to assess the interpretability, explainability, cybersecurity, and safety of ongoing updates. From the perspective of HTA agencies, we highlight that transparency, professional and patient acceptance, ethical issues, and organizational changes are required for the implementation of these devices. Economic assessments of AI should rely on a robust methodology (business impact or health economic models) to provide decision-makers with more reliable evidence.

Abbreviations: AI, Artificial intelligence; AI-based MD, Artificial intelligence-based medical device; AUC, Area under the curve; MD, Medical device; ML, Machine learning; SaMD, Software as a Medical Device; ITFoC, Information Technology: The Future of Cancer; Bfarm, Bundesinstitut Fur Arzneimittel und Medizinprodukte German Ministry of Health; CADTH, Canadian Agency for Drugs and Technologies in Health; FDA,, Food and Drug administration; HAS, Haute Autorité de Santé; NHS, National Health Service; NICE, National Institute for Health & Care Excellence; NIPH, Norwegian Institute of Public Health; INAHTA, International Network of Agencies for Health Technology Assessment; EUnetHTA, European network for Health Technology Assessment Joint Action.

* Corresponding author at: Groupe de Recherche et d'accueil en Droit et Economie de la Santé (GRADES) Department, University Paris-Saclay, Orsay, France.
E-mail addresses: line.farah1@gmail.com, l.farah@hopital-foch.com (L. Farah).

https://doi.org/10.1016/j.artmed.2023.102547
Received 31 May 2022; Received in revised form 28 March 2023; Accepted 4 April 2023
Available online 23 April 2023
0933-3657/© 2023 The Author(s). Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Conclusion: Currently, AI studies are insufficient to cover HTA prerequisites. HTA processes also need to be adapted because they do not consider the important specificities of AI-based MDs. Specific HTA workflows and accurate assessment tools should be designed to standardise evaluations, generate reliable evidence, and create confidence.

1. Introduction

The development of artificial intelligence (AI) in healthcare is growing exponentially. In the past 15 years, the number of articles dealing with medical AI has increased >60-fold, from 203 in 2005 to 12,563 in 2020 [1]. The implementation of AI in daily practice and its evaluation are currently of great interest to patients, healthcare professionals, and policymakers [2]. AI algorithms are particularly relevant in diagnostic, prognostic, and precision medicine, either in combination with medical devices (MDs) or as MDs by themselves. The term "AI-based MDs" denotes the use of computer technology to perform human tasks with or without health technologies. The Food and Drug Administration (FDA) classifies AI-based MDs as "Software as a Medical Device" (SaMD) when they are intended to treat, diagnose, cure, mitigate, or prevent disease or other conditions [3]. These AI-based MDs assist humans in completing tasks, but do not completely replace them because they are supervised by healthcare professionals.

However, no international consensus has been reached regarding the assessment of such health technologies. Specific reporting guidelines focusing on AI articles, particularly on methodological issues, have recently been suggested [4–10]. A high proportion of overlap can be identified in the aforementioned guidelines, highlighting the relative importance of some criteria; however, there is also an absence of consensus. However, such a consensus is essential for regulatory authorities and health technology assessment (HTA) agencies that face multiple challenges when assessing these AI technologies. On the one hand, regulatory authorities provide market authorization and certification to ensure that AI-based medical devices conform to legal requirements. By contrast, health technology assessment (HTA) agencies make recommendations for AI-based medical devices that can be financed or reimbursed by the healthcare system [11].

These AI-based MDs complexities make it difficult for healthcare professionals and patients to trust them. Similarly, HTA bodies need to be cautious about the ethical and regulatory implications of AI that could be considered barriers to the deployment of these technologies [11,12]. Because AI-based MDs differ from other health technologies, Dzobo et al. recommended regulating AI with specific legislation [13]. Uncertainties related to AI decision strategies and outcomes increase the difficulty of regulating these technologies [14]. Regulatory authorities, HTA bodies, and health policymakers are facing new challenges related to the new level of complexity in evaluating and delivering approval of AI-based MDs [2]; issues for the assessment of these devices relate to data generation, real-world usage, and undeveloped regulatory processes [15]. As highlighted by several authors, there is a growing need for specific health technology assessments of AI-based technologies. [2,16–18] However, these articles did not suggest a methodology to assess the quality of studies on AI-based MDs to evaluate the possibility to manage an HTA process. Therefore, in the present study, we propose a standardised methodological assessment method adapted for AI-based MDs intended to undergo HTA.

Therefore, from the payer's perspective, it has been suggested that a specific HTA process for AI should be designed, consisting of a multidisciplinary process that summarises data collected using a systematic, unbiased, transparent, and robust methodology [19,20]. Usually, HTA core models combine several items, including health problems and current use of the technology, description and technical characteristics of the technology, safety and clinical assessment, economic evaluation, and ethical, organizational, social, and legal aspects [20]. Although the aforementioned reporting guidelines represent an important first step in assessing the quality of the development of AI-based MDs and guaranteeing reliable AI tools, they are insufficient for designing robust and validated studies involving AI-based MDs in an HTA context [17]. Therefore, some regulatory and HTA bodies share strategies and recommendations on this topic. For example, the FDA has recently published an "Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan" [3,21–23]. The French Haute Autorité de Santé (HTA) agency suggested a list of 42 items divided into four categories to describe the technical characteristics of AI-based MDs [24]. In the United Kingdom, the National Institute for Health and Care Excellence (NICE) has proposed a guide for good practices in digital and data-driven health technologies, including AI [25]. Finally, the South Korean Ministry of Food and Drug Safety proposed guidelines to evaluate the clinical efficacy of MDs using AI [25].

Nevertheless, although European and American HTA bodies share approval decisions for AI-based MDs [26–28], specific studies presented to regulatory agencies are not systematically and publicly available. Therefore, we could not easily analyse successful methodologies that led to market access (e.g. CE marking in Europe and FDA approval in the USA). Therefore, it would be interesting to analyse whether the information in publicly available published studies corresponds to general HTA items related to MDs and specific items concerning AI. Whether the available studies meet the expectations and needs of current guidelines for HTA agencies is of particular interest.

The present study aimed to investigate, through a systematic literature review, whether the available guidelines apply to clinical studies of AI-based MDs and whether these generate sufficient and reliable evidence for the HTA process. The scope of this study was restricted to AI that could be assessed as MD, that is, in diagnosis, prognosis, screening, prevention, or treatment.

2. Materials and methods

2.1. Methods for searching for and selecting articles

Our review addresses the following questions: What are the study designs of current clinical studies on AI-based MDs, and are these studies of sufficient quality to support the relevant HTA processes?(2)(3)

To address this, we created a three-step study protocol (see Supplementary File 1) following three steps (Fig. 1).

(1) Part 1: a systematic literature review to identify articles assessing AI-based MDs
(2) Part 2: Identification of the critical criteria for the quality and HTA assessment of AI-based MDs
(3) Part 3: Evaluation of quality and HTA scores to assess whether studies generated reliable evidence for the HTA process.

2.2. Part 1: Study identification and search strategy

Comprehensive research was performed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (see PRISMA checklist in Supplementary File 1) reporting checklist [29] using search terms presented in the study protocol (Table 1). The databases used were PubMed (ncbi.nlm.nih.gov), Embase (Embase.com), the Cochrane Library (cochrane.org/fr/evidence), and HTA agency websites. This research was limited to the English and French languages. As Mesko et al. showed that the number of clinical studies has greatly increased

since 2016 [1], we retrieved studies published in the five years from 1 January 2016 to 31 December 2021. We did not find any AI-based MD assessments conducted by a health technology assessment agency before 2016. The first published HTA agency evaluation of an AI-based MD was in April 2018 for the evaluation of IDx-DR by the FDA, and as we wanted to evaluate the studies which could be analysed by HTA agencies, we consequently considered the time required for the development of these technologies (around two years), which corresponds to 2016, with an increasing publication number of AI articles published.

Additional sources were retrieved manually from the HTA body websites (Bfarm, CADTH, FDA, HAS, NHS, NICE, NIPH, INAHTA, and EUnetHTA) and gray literature.

2.2.1. Inclusion and exclusion criteria

Articles satisfying the following inclusion criteria were selected: original research articles published in peer-reviewed journals, and assessments of AI-based MDs. All types of clinical assessment studies were eligible for inclusion, including clinical evaluations or validations, external validations, economic evaluations, feasibility studies, randomized clinical trials, retrospective or prospective studies, and single-centre or multicentre studies.

Editorials, letters, comments, newspaper articles, and posters were excluded. Articles related to AI development methods (instead of assessments) were excluded, as were articles that presented animal experiments. Patients who failed to meet at least one of the eligibility criteria were excluded.

2.2.2. Study selection

The studies of interest were selected by four reviewers (LF, JDS, PN, and TM). After removing duplicates, four reviewers independently screened the abstracts to select eligible studies. Full-text reports were analysed for eligibility by the four reviewers. A fifth reviewer (NM) resolved the possible discrepancies highlighted during the selection process if a consensus was not reached. Third-party adjudication for dispute resolution was used for data extraction. An extraction database was used to list the selected studies that met the inclusion criteria and ensure that all eligible studies were included.

Table 1
Search strategy and MeSH terms used for our systematic literature review.

Search strategy
((Artificial intelligence[Title/Abstract]) OR (Machine learning[Title/Abstract]) OR (Artificial neural network[Title/Abstract]) OR (Support vector machine[Title/Abstract]) OR (SVM[Title/Abstract]) OR (CNN[Title/Abstract]) OR (RNN[Title/Abstract]) OR (LSTM[Title/Abstract]) OR (ResNet[Title/Abstract]) OR (DenseNet[Title/Abstract]) OR (Unet[Title/Abstract]) OR (DNN[Title/Abstract]) OR (Neural network*[Title/Abstract]) OR (Convolutional network*[Title/Abstract]) OR (Deep learn*[Title/Abstract]))
AND ((“Technology Assessment, Biomedical”[MeSH] OR “clinical evaluation”[All Fields] OR “Program Evaluation/methods”[MeSH] OR “Research Design”[MeSH] OR “Biomedical Research/methods”[MeSH] OR “Biomedical Research/standards”[MeSH] OR “Clinical Competence”[Mesh] OR “Decision Making”[MeSH] OR “Device Approval”[All Fields] OR “Diagnostic Test Approval”[MeSH] OR “Health Policy”[Mesh] OR “global health”[MeSH Terms] OR “Medical Device Legislation”[Mesh] OR “Ethics, Medical”[Mesh] OR “United States Food and Drug Administration”[Mesh] OR “Checklist”[Title/Abstract])
OR (“Biomedical Research/economics”[MeSH] OR “Evidence-Based Medicine/economics”[MeSH] OR “Costs and Cost Analysis”[MeSH] OR “Artificial Intelligence/economics”[MAJR] OR “Cost Savings”[Mesh] OR “Health Care Costs”[Mesh] OR “Economics” [MeSH]) OR (methodology[All Fields] AND “Humans”[Mesh] AND “Research Design”[Mesh]) OR (“Reproducibility of Results”[MeSH] AND (“clinical evaluation”[Title/Abstract] OR “Program Evaluation/methods”[MeSH])))
AND (English [Language] OR French [Language])
Filters: from 2016 to 2021

2.2.3. Data extraction

The following items were extracted from the articles selected by four analysts (LF, JDS, PN, and TM).

- General characteristics of the studies (authors, country, publication date, journal, study objectives, and assessment methodology)
- Characteristics of the AI technology (type, algorithm purpose)
- Dataset: Target population, categories (training, validation, test, data augmentation), quality and quantity
- Algorithm performance and validation

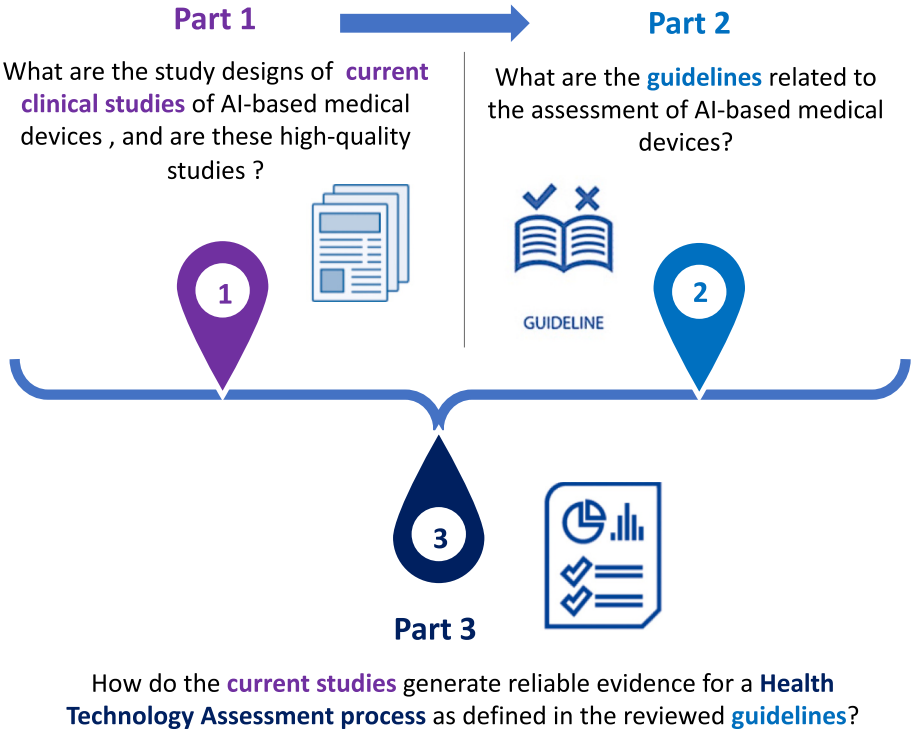

Fig. 1. Overall methodology and objectives of the article from the systematic literature review to the studies analysed for an HTA process.

- Reproducibility, code availability, explainability of the algorithm
- Evidence generation:1) outcomes (technological, clinical, economical, and side effects); 2) comparators/gold standards
- Ethical, legal, or social concerns

2.3. Part 2: Requirements of HTA bodies

We used the HTA Core Model® published by the European network for Health Technology Assessment (EUnetHTA) as a reference [30]. The HTA Core Model® is a methodological framework that assesses health technologies using a standardised method through nine domains: (1) health problems and current use of the technology, (2) description and technical characteristics of the technology, (3) safety, (4) clinical effectiveness, (5) costs and economic evaluation, (6) ethical analysis, (7) organizational aspects, (8) patients and social aspects, and (9) legal aspects.

Because of the uncertainty related to AI decision strategies and outcomes, data issues, and undeveloped regulatory processes, HTA agencies are facing various challenges to evaluate technologies and deliver approval [13–15]. The regulation of AI with specific legislation is needed. Therefore, we assessed articles not only for the HTA criteria but also for specific criteria described in the Part 3 to assess AI quality.

Regarding the HTA Core Model®, we searched each article for items in seven out of nine HTA Core Model domains (1 to 7), awarding 1 point if the item was present and 0 if not. As the items in domains 8 and 9 were outside the scope of our review, we did not search for these items, and they were not included in the scoring. The process for rating was double rating by LF and PN, and the mean score for each article was calculated.

2.4. Part 3: Quality assessment of the selected studies as compared to the HTA assessment

To assess the relationship between the quality and exhaustibility of the selected AI-based MD studies (Part 1) and the requirements of the HTA evaluation process (Part 2), we created two score tables, one for the AI quality assessment (Table 2) and one for the HTA domain evaluation (Table 3).

For AI quality assessment, we selected seven guidelines and checklists that could be used to assess AI studies. The results of the seven guidelines and checklists are summarised in the study protocol available in Supplementary file 1. Several criteria such as performance, data, assessment in clinical practice, reproducibility, and design were highlighted in each guideline. We noted that the MI-CLAIM checklist is the most complete, reliable, and general checklist that includes all the required criteria for AI assessment in healthcare [8]. Therefore, we used the MI-CLAIM checklist, which is a checklist of items suggesting a minimum set of data to enable assessment of clinical impact, build homogeneous levels of transparency, and assess the design process of AI-based MD clinical studies [8], to assess the articles in our study.

Each article was assessed using 21 items adapted from the MI-CLAIM checklist (Table 2). The items were grouped into six categories: study design (Category 1), data and optimisation (Categories 2 and 3), model performance (Category 4), model examination and assessment in clinical practice (Category 5), and reproducibility (Category 6). To adapt the checklist to our analysis needs, in categories 2 and 3, we listed items for which we were looking for quality technical and/or clinical data". In category 4, we added the two metrics summarised in item 15 (clinical and economic indicators) in addition to the technical evaluation criteria

Table 2
List of the AI criteria selected after analysis of the guidelines and adapted from items of the MI-CLAIM [8] checklist used for the AI quality assessment score table.

	Study design (category 1)
1	The clinical problem in which the model will be employed is clearly detailed in the paper.
2	The research question is clearly stated.
3	The characteristics of the cohorts (training and test sets) are detailed in the text. (0.5 point for training information / 0.5 point for test information).
4	The cohorts (training and test sets) are shown to be representative of real-world clinical settings.
5	The state-of-the-art solution used as a baseline for comparison has been identified and detailed.
6	Is there a comparator in the study?
7	If yes, is the comparator considered as a gold standard?

	Data and optimization (categories 2, 3)
8	The origin of the data is described, and the original format is detailed in the paper.
9	Data (technical and clinical) quality before it is applied to the proposed model is described.
10	The independence between training and test sets has been mentioned in the paper.
11	Data quantity is detailed and justified.
12	Targeted population is defined.
13	Is the input data type mentioned (structured or unstructured)?

	Model performance (category 4)
14	The primary metric selected to evaluate algorithm performance (e.g.: Area Under the Curve AUC, F-score, etc.) has been clearly stated (even in a previous study).
15	Summary of clinical and/or economic Key performance indicators (KPIs): The primary metric selected to evaluate the clinical and/or economic benefit is described.
16	The performance comparison between baseline and proposed model is presented with the appropriate statistical significance.

	Model examination/ assessment in clinical practice (category 5)
17	Explainability: Are the outputs of the algorithm clinically intelligible? Is the algorithm explainable?
18	Has use of the algorithm been shown to fit into and/or complement current clinical workflows?
19	Does the study answer the question about the possible patient harm or side effects (ex: does this cause a delay in diagnosis?) that could be caused by the algorithm?
20	Does use of the algorithm raise ethical, legal, or social concerns?

	Reproducibility (category 6)
21	Total score of (1) Reproducibility/ generalizability in clinical practice and (2) Code source / code availability or not mentioned in the study.

Table 3
Score attributed to articles for each HTA domain (excluding domains 8 and 9).

Author	(1) Health problem and current use of technology	(2) Description and technical characteristics of technology	(3) Safety	(4) Clinical effectiveness	(5) Costs and economic evaluation	(6) Ethical aspect	(7) Organizational aspects	Score of completion of HTA domains
Schreier, J. et al	1	1	1	1	0	0	1	71%
Ladefoged, C.N. et al	1	1	0	1	1	0	1	71%
Im, H.et al	1	1	0	1	1	0	1	71%
Ohta, Y. et al	1	1	1	1	0	0	1	71%
Tseng, A. et al	1	1	0	1	1	0	1	71%
Schwendicke, F. et al	1	1	0	1	1	0	1	71%
Nathan R. H. et al	1	1	0	1	1	0	1	71%
R. H. H. M. Philipsen et al	1	1	0	1	1	0	1	71%
Tufail, V Kapetanakis, et al	1	1	0	1	1	0	1	71%
Risa M. Wolf, et al	1	1	0	1	1	0	1	71%
Faes, L.	1	1	0	1	1	0	0	57%
Kemnitz, J. et al	1	1	1	1	0	0	0	57%
Zou, F.-W. et al	1	1	0	1	0	0	1	57%
Mergen, V. et al	1	1	0	1	1	0	0	57%
Yang, S. et al	1	1	0	1	0	0	1	57%
Kanagasingam, Yogesan et al	1	1	0	1	0	0	1	57%
Böttcher, Benjamin et al	1	1	0	1	1	0	0	57%
Stuckey, Thomas D et al	1	1	1	1	0	0	0	57%
Yuwei Liu et al	1	1	0	1	0	0	1	57%
Grzybowski, A. et al	1	1	0	1	0	0	1	57%
Medina, R. et al	1	1	1	1	0	0	0	57%
Jefferies, J.L. et al	1	1	0	1	0	0	1	57%
Yeh, Eric et al	1	1	0	1	0	0	1	57%
McLouth, J. et al	1	1	0	1	0	0	1	57%
Joo, B. et al	1	1	0	1	0	0	1	57%
Wei, Yingting et al	1	1	0	1	0	0	1	57%
Yang, W.-H. et al	1	1	0	1	0	0	0	43%
Zabel, W. Jeffrey et al	1	1	0	1	0	0	0	43%
Kim, J.H. et al	1	1	0	1	0	0	0	43%
Ahn, Sang Hee et al	1	1	0	1	0	0	0	43%
Chen, C. et al	1	1	0	1	0	0	0	43%
Liu, Z. et al	1	1	0	1	0	0	0	43%
Dembrower, K. et al	1	1	0	1	0	0	0	43%
Benjamins, J.W. et al	1	1	0	1	0	0	0	43%
Brunenberg, E.J.L. et al	1	1	0	1	0	0	0	43%
Krause, Jonathan et al	1	1	0	1	0	0	0	43%
Ihlen, E.A.F. et al	1	1	0	1	0	0	0	43%
Sun, Chao et al	1	1	0	1	0	0	0	43%
Wu, Yijun et al	1	1	0	1	0	0	0	43%

Connolly, P. et al	1	1	0	1	0	0	0	43%
Martins Jarnalo, C. O. et al	1	1	0	1	0	0	0	43%
Xie, Yuchen et al	1	1	0	1	0	0	0	43%
Perkuhn, Michael et al	1	1	0	1	0	0	0	43%
Choi, Min Seo et al	1	1	0	1	0	0	0	43%
Brenton, Lisa et al	1	1	0	1	0	0	0	43%
Mehralivand, Sherif et al	1	1	0	1	0	0	0	43%
Winkel, D.J. et al	1	1	0	1	0	0	0	43%
Pennig, Lenhard et al	1	1	0	1	0	0	0	43%
Rudie, J.D. et al	1	1	0	1	0	0	0	43%
Potash, Eric et al	1	1	0	1	0	0	0	43%
Balidis, M. et al	1	1	0	1	0	0	0	43%
Cikes, Maja et al	1	1	0	1	0	0	0	43%
Attia, Zachi I et al	1	1	0	1	0	0	0	43%
Cesaretti, Manuela et al	1	1	0	1	0	0	0	43%
Shah, Payal et al	1	1	0	1	0	0	0	43%
Zech, John R et al	1	1	0	1	0	0	0	43%

Legend: score 1 if the criterion is described in the study; score 0 if the criterion is not described.
In dark green: studies scored > 70 % for the seven HTA domains
In light green: studies scored between 50 % and 70 % for the seven HTA domains
In yellow: studies scored < 50 % for the seven HTA domains

(item 14). In Category 6 (reproducibility), we combined the two criteria of code generalisability and availability into one criterion (Item 21).

Each category includes several items and questions presented in the study protocol (Table 2). For each selected study, each item was rated as 0 if the item was absent from the article, 0.5 points if the item was partially completed, or 1 if the criterion was fully completed.

2.5. Statistical analysis

We conducted a linear regression with the overall HTA score, explanatory variable impact factor, and AI score (i.e. the overall MI-CLAIM score). Journal impact factors were retrieved from each journal in 2021. We also performed linear regression with the AI (MI-CLAIM) score, explanatory variables' impact factors, and HTA scores. Statistical analysis was performed using R software (version 4.1.1; 2020) - R Foundation for Statistical Computing, Vienna, Austria. Following the statistical healthcare standards mentioned by several authors [31,32], we used 95 % confidence intervals, an alpha risk inferior to 5 %, and a *p*-value <5 %.

As highlighted in the literature, several authors showed the impact factor (IF) as a reasonable indicator of quality. Therefore, we analysed the correlation between IF and quality scores of the articles [33–36].

We performed linear regression with the outcome HTA Score, explanatory variable Impact Factor and AI Score (MI-CLAIM). We performed linear regression with the AI Score outcome, explanatory variables' impact factors, and HTA scores.

For linear regression, the candidate adjustment variables were introduced into the Least Absolute Shrinkage and Selection Operation (LASSO) penalized regression model [37]. To select the variables, we used a LASSO-type model between the variable to be explained and the explanatory variables. The penalty coefficient (λ lambda) was chosen so that it provided an estimation error of less than one standard deviation of the minimum error obtained by cross-validation 10 times, while being as parsimonious as possible. We used the largest value of λ for which the cross-validation error is within 1 standard error of the minimum cross-validation error. If using this parameter, the number of nonzero coefficients was less than the maximum number of covariates, this parameter was maintained. Otherwise, we chose the highest value of λ which provides several coefficients equal to the maximum number of covariates. No variable had a coefficient different from 0 with this lambda λ coefficient.

As the numbers compared were small, a non-parametric test was carried out using the Kruskal-Wallis test for AI and HTA scores and impact factors. Fisher's exact test with an alpha risk of 5 % was conducted for medical specialty variables.

We conducted a univariate analysis of the HTA score with an alpha risk of 5 % by adjusting for the impact factor and HTA score. We conducted a multivariate analysis of the AI score with an alpha risk of 5 % by adjusting for the impact factor and HTA score.

As the numbers compared were small, a non-parametric test was performed using the Kruskal-Wallis test for AI and HTA scores, publication date, and impact factor. Fisher's exact test with an alpha risk of 5 % was used for the medical specialty variable. There were no missing data in this dataset.

3. Results

3.1. Selection

The study selection retrieved 5578 records in total (see PRISMA Flowchart in Fig. 2) with 1035 duplicates, 4543 screened in title-abstract, 4450 excluded in title-abstract, 37 excluded in full text, and 56 finally included in the review [38–93]. The main reason for the exclusion was the focus on the development rather than the assessment of AI-based MDs.

All results and data extracted from the 56 articles are summarised in Supplementary file 2. Of the 56 articles, 71 % ($n = 40$) of the selected studies were published after 2020. The authors' country affiliations were distributed across Europe (43 %, $n = 24$), Asia (28.5 %, $n = 16$), North America (25 %, $n = 14$), and Australia (3.5 %, n = 2). >50 % of

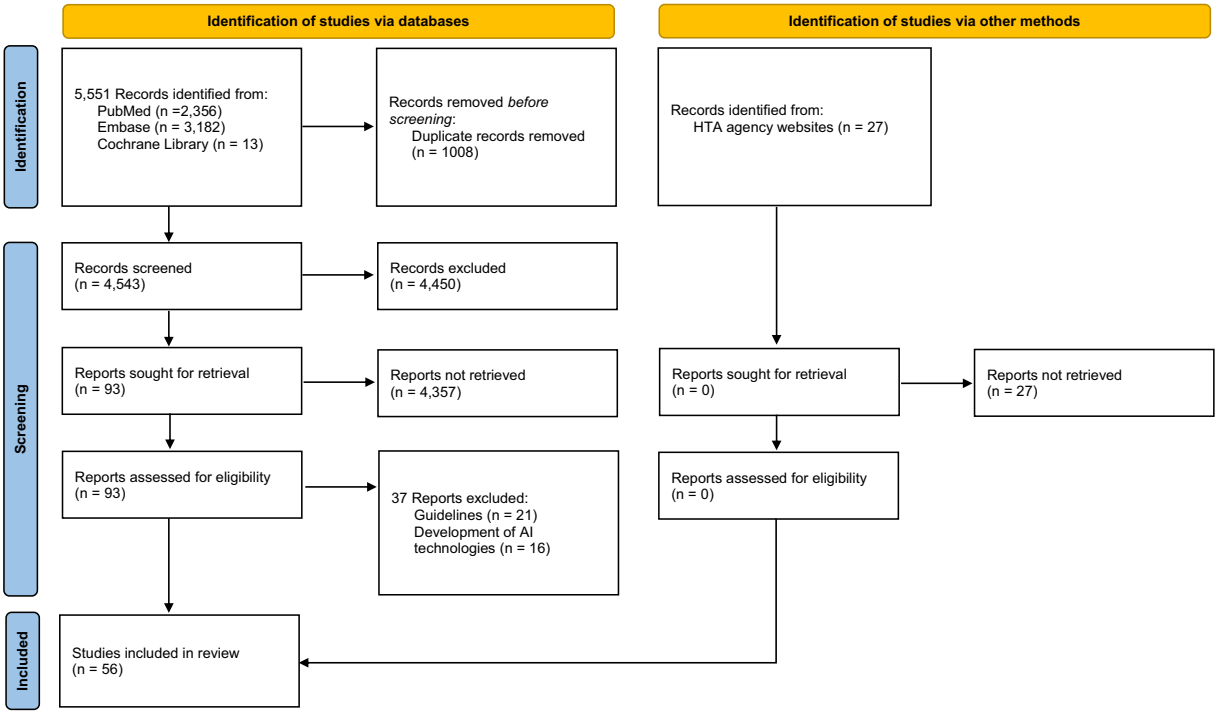

Fig. 2. PRISMA flow chart for study selection process.

the studies focused on radiology and imaging. Regarding study design, 89 % ($n = 50$) were retrospective studies, whereas only 11 % ($n = 6$) were prospective studies. Only two studies were randomized clinical trials [75,85]. Of the selected studies, 18 % ($n = 10$) were multicentre.

3.2. Results of the AI quality assessment

The use of the MI-CLAIM checklist to assess the studies' AI quality resulted in a mean score of 67 %. Of the 56 articles, 32 % ($n = 18$) were rated as having an AI quality score over or equal to 70 %, 50 % ($n = 28$) had a score between 50 % and 70 %, and 18 % (n = 10) had a score below 50 %. The study design category scored the highest (82 %). The data/optimization and model performance categories were completed in 69 % and 66 % of the cases, respectively, whereas the model examination/assessment in clinical practice and reproducibility categories had the lowest scores (Fig. 3). Detailed results are presented in Supplementary File 3.

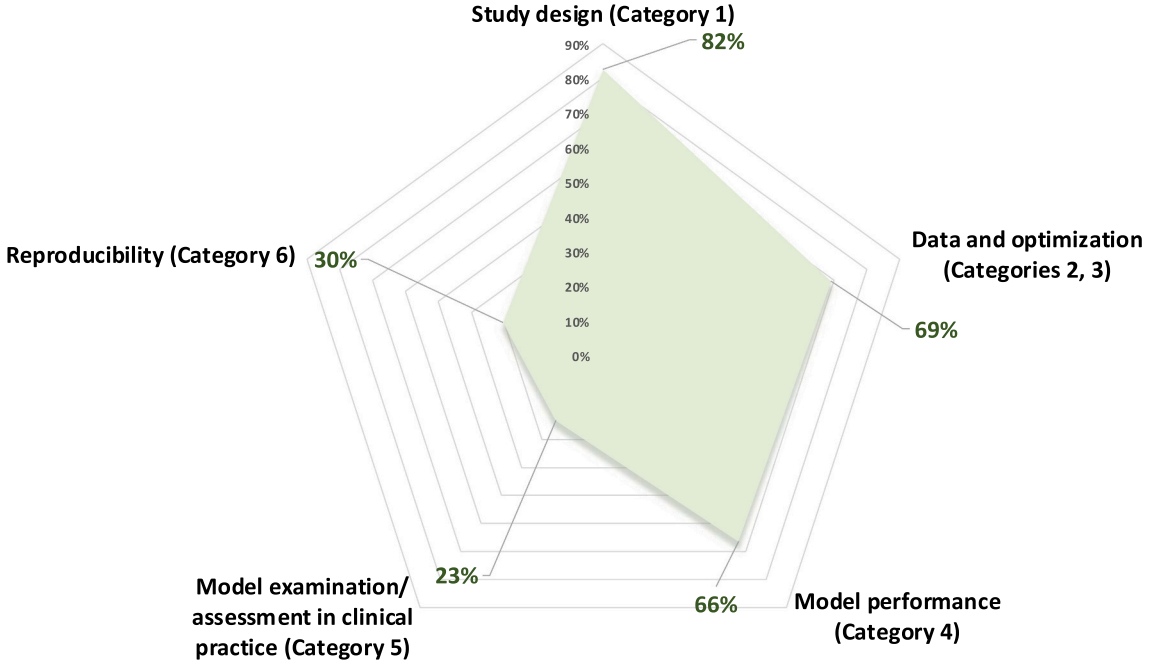

Fig. 3. Qualitative assessment of articles based on the criteria of the guidelines reviewed in Part 1 and adapted from the MI-CLAIM checklist
Legend: Percentages show proportion of articles completing each category.

3.3. Results of the HTA domain evaluation

For the HTA Core Model® items, the mean score was 48 % for all seven domains. We found that 18 % of the studies scored >70 % for the seven HTA domains, 29 % of studies obtained a scored between 50 % and 70 %, and 53 % scored <50 % (Table 3).

Three HTA domains (health problems/current use, description of technology, and clinical effectiveness) were systematically described; however, only 9 % ($n = 5$) of the articles evaluated the safety of the device. Economic aspects were discussed by only 20 % ($n = 11$) of the articles, and no article mentioned the analysis of the ethical criteria related to the assessment of AI-based MDs (Fig. 4).

3.4. Univariate analysis

The mean AI (MI-CLAIM) score and medical specialty did not significantly differ according to the HTA score (respectively $p = 0.19$, $p = 0.475$) (Table 4). In contrast, the average rank of the impact factors differed significantly according to the HTA score ($p = 0.046$). On the one side, the AI score and medical specialty were not modified by the HTA score even if it was a low or a high score. On the other hand, the journal impact factor seemed to influence the HTA score assessment.

3.5. Multivariate analysis

As shown in the univariate analysis, there was a statistically significant relationship between the AI (MI-CLAIM) score and impact factors. When the impact factor increased by one unit (for instance, from 6 to 7), the average AI score increased by 0.00490 ($p \leq 0.01$ (Supplementary File 4). There was no statistically significant difference in the AI score based on the HTA Score ($p = 0.19$).

4. Discussion

To the best of our knowledge, the present work is the first to perform a literature review of published AI-based MD assessment studies and compare the items present in these studies to specific HTA criteria related to AI-based MDs to assess whether study quality corresponds with HTA requirements. We observed that, in general, studies related to AI-based MDs do not sufficiently fulfil HTA criteria, with widely heterogeneous completion rates of the included HTA Core Model® items and a mean completion rate of 52 %. Despite the promise of AI-based MDs in improving clinical and economic outcomes, the assessment of the actual value of AI technologies in real-world clinical practice remains an important challenge.

The main question related to the assessment of AI in healthcare is, "which unmet need is the AI-based MD going to solve?" [94]. All studies in our review responded to this question. However, the clinical evaluation of AI-based MD and its methodology should be highly reliable, consistent, and sufficiently robust to handle situations such as missing or false data. Our results suggest that clinical studies on AI-based MDs have several limitations, including a lack of adapted, robust, transparent, and complete evidence.

Our results appear to be consistent with the literature in terms of the analysis of studies from the perspective of regulatory authorities and the HTA body. For instance, in the USA, Wu et al. assessed AI devices approved by the FDA between 2015 and 2020 [95] and concluded that most evaluations were retrospective studies, as was also observed in our present study, in which almost 80 % of the studies were retrospective. Although 80 of the 124 AI-based MDs approved in both the USA and Europe were first approved in Europe according to Muehlematter et al., the authors highlighted the difficulties of studying CE-marked MDs in Europe (unlike in the USA, there is no publicly available register of approved MDs because of the confidentiality of information delivered to HTA agencies and the decentralised pathway of CE marking) [29]. Therefore, one of the strengths of our article is that we propose both quantitative assessment tools for and qualitative analysis of the evidence available in AI-based MD studies.

4.1. Assessment of the quality of datasets required for the HTA process

To guarantee a robust methodology for the evaluation of AI, studies must include a justification of the sample size and an assessment of data quality to ensure the replicability of results [96]. In our review, >70 %

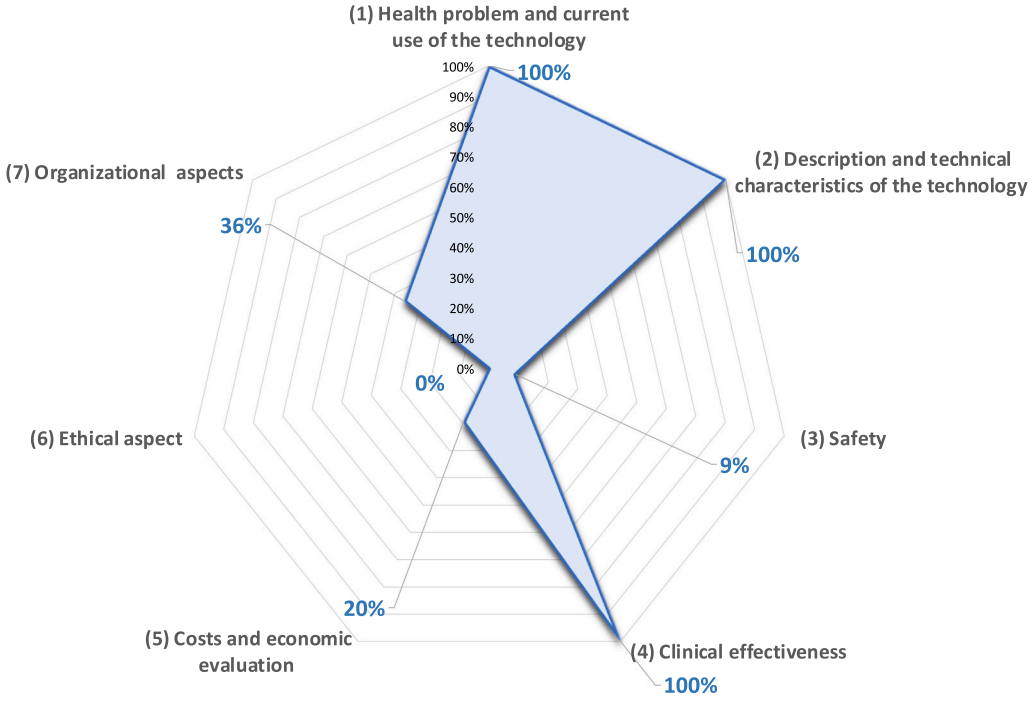

Fig. 4. Proportion of articles completing each of the HTA Core Model® selected domains
Legend: Percentages show proportion of articles completing each category.

Table 4
Univariate analysis of the distribution of the HTA score related to the AI score, impact factor and medical specialty.

HTA Score		HTA Score 0.43 ($n = 30$)	HTA Score 0.57 (n = 16)	HTA Score 0.71 ($n = 10$)	n	p	Test
AI Score - MI-CLAIM, median [Q25–75]		0.600 [0.500; 0.700]	0.625 [0.550; 0.700]	0.650 [0.612; 0.756]	56	0.17	Kruskal-Wallis
Impact Factor, mean (standard deviation)		5.98 (9.61)	3.68 (1.71)	7.55 (6.62)	56	**0.046***	Kruskal-Wallis
Medical specialty, n	Radiology	13 (43 %)	5 (31 %)	4 (40 %)	22	0.47	Fisher
	Other	7 (23 %)	7 (44 %)	1 (10 %)	15	–	–
	Ophthalmology - diabetes	4 (13 %)	2 (12 %)	2 (20 %)	8	–	–
	Cardiovascular	3 (10 %)	0 (0 %)	2 (20 %)	5	–	–
	Oncology	1 (3.3 %)	2 (12 %)	1 (10 %)	4	–	–
	Uro-Nephrology	2 (6.7 %)	0 (0 %)	0 (0 %)	2	–	–

Legend: * = significative results ($p < 0.05$).

of the articles did not justify their sample size. Navarro et al. highlighted the poor methodological quality of most studies on machine learning prediction models, revealing a high risk of bias related to small study sizes, inappropriate handling of missing data, and failure to deal with overfitting [97]. High-quality datasets are required because the output data can only be trusted if the inputs are reliable. Although almost all the selected studies in the present review described the origin of the dataset, more than one-third did not describe the technical and clinical data quality assessment applied to their dataset before its use within the proposed AI model. Several AI applications in the reported studies are not exploitable in clinical practice because they are trained with complete datasets, but may not be transposable to the real-world", where data can be of poor quality or incomplete [2]. However, the results of an algorithm trained on large datasets do not imply that they are generalisable.

In addition, the code for the AI algorithm was available in only a few studies. Similarly, in the USA, Amann et al. highlighted in 2020 that, because of intellectual property issues, less information is available on the methodologies for FDA-approved AI models than on those for open-source solutions [98]. FDA-approved AI technologies also use smaller datasets because public datasets are available only to academic and non-commercial entities [98]. According to Reston et al., universal test sets are required to compare regulatory bodies across AI-based MD manufacturers [99]. The absence of universal test sets could lead to bias in the test sets, patient populations, or even in the way data were acquired.

4.2. Recommendations for adapting HTA criteria to AI-based MDs

Although international reporting guidelines for AI technologies have been developed (SPIRIT-AI [4], CONSORT-AI [5], STARD-AI [6], CLAIM [7], MI-CLAIM [8], PROBAST-ML [9]), there remains a lack of consensus regarding the assessment and evaluation of AI-based MDs.

4.2.1. Recommendations for regulatory authorities

In Europe, the introduction of the Medical Device Regulation (EU 2017/745 or MDR) has highlighted the need for more in-depth clinical data to prove safety and performance claims, including tighter equivalency standards.

Consequently, we suggest the following recommendations from the perspective of regulatory authorities, such as the US FDA or European notified bodies:

- Requirements for quality management systems and post-market surveillance systems should be reinforced [100]. This is particularly true for AI-based MDs that require regular audits, ongoing monitoring, and reporting systems to evaluate the safety, quality, transparency, and ethical factors of AI-based services [101].
- The FDA also addressed the need for iterative modifications by publishing a discussion paper in 2019 proposing a regulatory framework for AI applications that can adapt to ongoing changes with new good machine learning practices, focusing on transparency and continued analytical and clinical validity [23].
- Among the challenges posed by AI, transparency, interpretability, and explainability are important, because they can influence trust in these technologies [102]. During the assessment process, these criteria should be considered to improve the trust of patients and healthcare professionals in AI. However, in the present work, a majority of the assessed studies did not describe how the AI algorithm works. The concept of explainability, which allows for an understanding of why the AI technology came up with a conclusion, is highly important for AI-based MDs [16,23].

4.2.2. Recommendations for HTA agencies

A consensus on the specific HTA criteria required to evaluate AI is required by HTA agencies that face multiple challenges when assessing these technologies [2,103–105]. Our systematic review highlights that specific HTA workflows and assessment tools should be used to standardise the evaluation of AI-based MDs. Therefore, we propose the following recommendations:

- During the HTA process of AI-based MDs, the criteria of transparency, interpretability, explainability, ethics, human-AI interaction, and organizational impact should be considered, in addition to the usual HTA criteria for health technologies (such as effectiveness and safety).
- Economic assessments of AI should rely on a more robust methodology, such as business impact models or specific health economic models, to provide stakeholders in healthcare decision-making with more reliable evidence.
- The quality of data management (collection, storage, privacy, and governance) should be considered as a criterion of AI-based MDs quality in the HTA evaluation process of AI-based MDs, but also for post-market surveillance, especially for evolutive algorithms.

The existing HTA frameworks are not specifically designed to assess AI-based MDs; however, a core HTA can be used as the basis for producing specific HTA reports. Our review suggests that existing HTA frameworks must be adjusted to appropriately assess AIMDs by including interpretability [106], explainability [102], cybersecurity [83], the clinical safety of algorithm updates, interoperability [107,108], professional and patient acceptance, and ethical and legal issues [109]. However, accurate criteria for assessing AI-based MDs in the HTA evaluation process are still lacking. For example, considering the organizational impact domain, as AI could lead to a redistribution of work among healthcare professionals, the assessment of AI should include the organizational changes required for its implementation in routine practice. Alami et al. highlighted the importance of studying organizational impact and readiness to integrate AI into healthcare delivery. [110] The organizational impacts could be related to a shorter delay in diagnosis, allowing physicians to introduce earlier treatment. It

is also related to the need for training new stakeholders who could follow the alerts generated by AI-based MD in the long-term follow-up of chronic diseases such as diabetes. At this stage, more studies are needed to address the lack of research on organizational issues raised by the integration of AI into clinical routines. [110]. However, our review highlighted that organizational impact was assessed in only 38 % of the studies ($n = 22$). In addition, AI is often seen as a potential way to reduce costs, rather than as a tool that requires additional resources to ensure proper functioning. However, this aspect was not properly assessed in the reviewed articles because only 21 % ($n = 12$) of the articles addressed this point.

4.3. Recommendations on AI-based MDs study design

Given that several criteria were largely unreported in the reviewed articles, future studies should address and monitor these issues. These criteria should be included as secondary endpoints in the study design. The aim was to standardise the HTA pathway and peer-reviewed journal process.

Regarding the HTA criteria, « safety » should be systematically assessed during the initial clinical study, even if the AI-based medical device appears to have a limited impact on the healthcare pathway. In addition, economic and organizational impact evaluations should be conducted. Ethical criteria must also be integrated into early evaluation studies to improve physicians' confidence in AI-based technologies [111,112]. Concerning the AI criteria, the "model assessment in clinical practice" and the "reproducibility" in other clinical environments should be assessed and clearly stated in the clinical studies to ensure a transparent and exhaustive evaluation process. The clear identification of these criteria as secondary endpoints in studies could also be highlighted during the review process of academic journals to meet higher-quality standards [113].

4.4. Limitations

Finally, our study has some limitations. We only included studies in English and French, which may have been regarded as biased. We limited our research to AI algorithms that were assessed as MDs or that could potentially be considered an MD without confirmation of MD status. In addition, the scores calculated for the quality assessment of the studies were based on items adapted from the MI-CLAIM checklist and HTA Core Model. As we did not find any specific HTA criteria or alternative HTA processes that could be used to evaluate AI-based MDs, we adapted a list of items to address the questions posed in our review. The criteria of the HTA Core Model are not completely suitable for the evaluation of clinical studies; this is particularly true for items in domains 8 and 9, which is why we excluded them, but also for items in domain 6 (ethics). Because HTA submissions are not usually mentioned in the objectives of the studies, this could be a limitation in the selection process of the systematic review. Another limitation of this study is the lack of access to studies assessed by HTA agencies, except for the article related to AI-based MD for diabetic retinopathy assessed by the FDA. Some authors have highlighted the importance of being cautious about journal impact factors because a correlation between IF and quality scores is not always found. [114]

5. Conclusion

This study focuses on the information that studies on AI-based MDs should contain to meet the expectations of HTA agencies and healthcare stakeholders. We found that AI studies are currently insufficient to meet all HTA expectations (safety, organizational impact, cost, and economic evaluation). Nevertheless, HTA evaluation criteria also need to be adapted, as they do not consider the important specificities of AI-based MDs [2,17–19] such as (1) the quality of clinical datasets on which the performance of the device relies and which are still of poor quality and not standardised, (2) interpretability and explainability [115], which drive user acceptability, (3) interoperability, and (4) reproducibility. Specific HTA workflows and accurate assessment tools should be designed to standardise the evaluation of AI-based MDs. Such improvements can shape value-based healthcare for AI by generating reliable evidence and creating confidence in health technologies.

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

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.artmed.2023.102547.

References

[1] Meskó B, Görög M. A short guide for medical professionals in the era of artificial intelligence. Npj Digit Med 2020;3:126. https://doi.org/10.1038/s41746-020-00333-z.
[2] Alami H, Lehoux P, Auclair Y, de Guise M, Gagnon M-P, Shaw J, et al. Artificial intelligence and health technology assessment: anticipating a new level of complexity. J Med Internet Res 2020;22:e17707. https://doi.org/10.2196/17707.
[3] Harvey HB, Gowda V. How the FDA regulates AI. Acad Radiol 2020;27:58–61. https://doi.org/10.1016/j.acra.2019.09.017.
[4] Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI, Group CONSORT-AIWorking, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 2020;26:1351–63. https://doi.org/10.1038/s41591-020-1037-7.
[5] Liu X, Cruz Rivera S, 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;26:1364–74. https://doi.org/10.1038/s41591-020-1034-x.
[6] Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, et al. Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open 2021;11:e047709. https://doi.org/10.1136/bmjopen-2020-047709.
[7] Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2020;2:e200029. https://doi.org/10.1148/ryai.2020200029.
[8] Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 2020;26:1320–4. https://doi.org/10.1038/s41591-020-1041-y.
[9] Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, 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:e048008. https://doi.org/10.1136/bmjopen-2020-048008.
[10] Tsopra R, Fernandez X, Luchinat C, Alberghina L, Lehrach H, Vanoni M, et al. A framework for validating AI in precision medicine: considerations from the european ITFoC consortium. BMC Med Inform Decis Mak 2021;21:274. https://doi.org/10.1186/s12911-021-01634-3.
[11] Ofori-Asenso R, Hallgreen CE, De Bruin ML. Improving interactions between health technology assessment bodies and regulatory agencies: a systematic review and cross-sectional survey on processes, progress, outcomes, and challenges. Front Med 2020:7.
[12] Bærøe K, Miyata-Sturm A, Henden E. How to achieve trustworthy artificial intelligence for health. Bull World Health Organ 2020;98:257–62. https://doi.org/10.2471/BLT.19.237289.
[13] Vayena E, Blasimme A, Cohen IG. Machine learning in medicine: addressing ethical challenges. PLoS Med 2018;15:e1002689. https://doi.org/10.1371/journal.pmed.1002689.
[14] Dzobo K, Adotey S, Thomford NE, Dzobo W. Integrating artificial and human intelligence: a partnership for responsible innovation in biomedical engineering and medicine. OMICS J Integr Biol 2020;24:247–63. https://doi.org/10.1089/omi.2019.0038.
[15] Zawati M, Lang M. What's in the Box?: uncertain accountability of machine learning applications in healthcare. Am J Bioeth 2020;20:37–40. https://doi.org/10.1080/15265161.2020.1820105.

[16] Gerke S, Babic B, Evgeniou T, Cohen IG. The need for a system view to regulate artificial intelligence/machine learning-based software as medical device. Npj Digit Med 2020;:3. https://doi.org/10.1038/s41746-020-0262-2.
[17] Hendrix N, Veenstra DL, Cheng M, Anderson NC, Verguet S. Assessing the economic value of clinical artificial intelligence: challenges and opportunities. Value Health J Int Soc Pharmacoeconomics Outcomes Res 2022;25:331–9. https://doi.org/10.1016/j.jval.2021.08.015.
[18] Bélisle-Pipon J-C, Couture V, Roy M-C, Ganache I, Goetghebeur M, Cohen IG. What makes artificial intelligence exceptional in health technology assessment? FrontArtif Intell 2021:4.
[19] Unsworth H, Wolfram V, Dillon B, Salmon M, Greaves F, Liu X, et al. Building an evidence standards framework for artificial intelligence-enabled digital health technologies. Lancet Digit Health 2022;4:e216–7. https://doi.org/10.1016/S2589-7500(22)00030-9.
[20] Kristensen FB, Husereau D, Huić M, Drummond M, Berger ML, Bond K, et al. Identifying the need for good practices in health technology assessment: summary of the ISPOR HTA Council working group report on good practices in HTA. Value Health 2019;22:13–20. https://doi.org/10.1016/j.jval.2018.08.010.
[21] HTA Core Model® – EUnetHTA n.d. https://www.eunethta.eu/hta-core-model/ (accessed October 20, 2021).
[22] Allen B. The role of the FDA in ensuring the safety and efficacy of artificial intelligence software and devices. J Am Coll Radiol 2019;16:208–10. https://doi.org/10.1016/j.jacr.2018.09.007.
[23] Artificial Intelligence and Machine Learning in Software as a Medical Device | FDA n.d. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device (accessed October 20, 2021).
[24] Software and AI as a Medical Device Change Programme. GOVUK n.d. https://www.gov.uk/government/publications/software-and-ai-as-a-medical-device-change-programme/software-and-ai-as-a-medical-device-change-programme (accessed October 20, 2021).
[25] Grille descriptive des fonctionnalités des dispositifs médicaux embarquant un système avec apprentissage automatique (intelligence artificielle). Haute Aut Santé n.d. https://www.has-sante.fr/jcms/p_3318028/fr/grille-descriptive-des-fonctionnalites-des-dispositifs-medicaux-embarquant-un-systeme-avec-apprentissage-automatique-intelligence-artificielle (accessed November 11, 2022).
[26] Regulations|Medical Devices|Our Works|Minisry of Food and Drug Safety n.d. https://www.mfds.go.kr/eng/brd/m_40/view.do?seq=72623&srchFr=&srchTo=&srchWord=&srchTp=&itm_seq_1=0&itm_seq_2=0&multi_itm_seq=0&company_cd=&company_nm=&page=1 (accessed October 20, 2021).
[27] Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 2020;3:118. https://doi.org/10.1038/s41746-020-00324-0.
[28] Health C for D and R. Artificial intelligence and machine learning (AI/ML)-enabled medical devices. FDA; 2021.
[29] Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health 2021;3:e195–203. https://doi.org/10.1016/S2589-7500(20)30292-2.
[30] Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, The PRISMA, et al. Statement: an updated guideline for reporting systematic reviews. BMJ 2020;2021:n71. https://doi.org/10.1136/bmj.n71.
[31] Lampe K, Mäkelä M, Garrido MV, Anttila H, Autti-Rämö I, Hicks NJ, et al. The HTA core model: a novel method for producing and reporting health technology assessments. Int J Technol Assess Health Care 2009;25(Suppl 2):9–20. https://doi.org/10.1017/S0266462309990638.
[32] Upshur RE. The ethics of alpha: reflections on statistics, evidence and values in medicine. Theor Med Bioeth 2001;22:565–76. https://doi.org/10.1023/a:1014462116530.
[33] Cucherat M, Laporte S. False positive results or what's the probability that a significant P-value indicates a true effect? Therapie 2017;72:421–6. https://doi.org/10.1016/j.therap.2016.09.021.
[34] Saha S, Saint S, Christakis DA. Impact factor: a valid measure of journal quality? J Med Libr Assoc 2003;91:42–6.
[35] Abi Jaoude J, Kouzy R, Rooney M, Thompson P, Patel R, Turner MC, et al. Impact factor and citation metrics in phase III cancer trials. Oncotarget 2021;12:1780–6. https://doi.org/10.18632/oncotarget.28044.
[36] Callaham M, Wears RL, Weber E. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. JAMA 2002;287:2847–50. https://doi.org/10.1001/jama.287.21.2847.
[37] Schneeweiss S, Eddings W, Glynn RJ, Patorno E, Rassen J, Franklin JM. Variable selection for confounding adjustment in high-dimensional covariate spaces when analyzing healthcare databases. Epidemiology 2017;28:237–48. https://doi.org/10.1097/EDE.0000000000000581.
[38] Yang W-H, Zheng B, Wu M-N, Zhu S-J, Fei F-Q, Weng M, et al. An evaluation system of fundus photograph-based intelligent diagnostic technology for diabetic retinopathy and applicability for research. Diabetes Ther 2019;10:1811–22. https://doi.org/10.1007/s13300-019-0652-0.
[39] Xie Y, Nguyen QD, Hamzah H, Lim G, Bellemo V, Gunasekeran DV, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. Lancet Digit Health 2020;2:e240–9. https://doi.org/10.1016/S2589-7500(20)30060-1.
[40] Faes L, Wagner SK, Fu DJ, Liu X, Korot E, Ledsam JR, et al. Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study. Lancet Digit Health 2019;1:e232–42. https://doi.org/10.1016/S2589-7500(19)30108-6.
[41] Schreier J, Genghi A, Laaksonen H, Morgas T, Haas B. Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam CT and CT. Radiother Oncol J Eur Soc Ther Radiol Oncol 2020;145:1–6. https://doi.org/10.1016/j.radonc.2019.11.021.
[42] Perkuhn M, Stavrinou P, Thiele F, Shakirin G, Mohan M, Garmpis D, et al. Clinical evaluation of a multiparametric deep learning model for glioblastoma segmentation using heterogeneous magnetic resonance imaging data from clinical routine. Invest Radiol 2018;53:647–54. https://doi.org/10.1097/RLI.0000000000000484.
[43] Choi MS, Choi BS, Chung SY, Kim N, Chun J, Kim YB, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiother Oncol J Eur Soc Ther Radiol Oncol 2020;153:139–45. https://doi.org/10.1016/j.radonc.2020.09.045.
[44] Zabel WJ, Conway JL, Gladwish A, Skliarenko J, Didiodato G, Goorts-Matthews L, et al. Clinical evaluation of deep learning and atlas-based auto-contouring of bladder and rectum for prostate radiation therapy. Pract Radiat Oncol 2021;11:e80–9. https://doi.org/10.1016/j.prro.2020.05.013.
[45] Kemnitz J, Baumgartner CF, Eckstein F, Chaudhari A, Ruhdorfer A, Wirth W, et al. Clinical evaluation of fully automated thigh muscle and adipose tissue segmentation using a U-net deep learning architecture in context of osteoarthritic knee pain. Magma N Y N 2020;33:483–93. https://doi.org/10.1007/s10334-019-00816-5.
[46] Brenton L, Waters MJ, Stanford T, Giglio S. Clinical evaluation of the APAS® Independence: automated imaging and interpretation of urine cultures using artificial intelligence with composite reference standard discrepant resolution. J Microbiol Methods 2020;177:106047. https://doi.org/10.1016/j.mimet.2020.106047.
[47] Kim JH, Kim JY, Kim GH, Kang D, Kim IJ, Seo J, et al. Clinical validation of a deep learning algorithm for detection of pneumonia on chest radiographs in emergency department patients with acute febrile respiratory illness. J Clin Med 2020;9:1–11. https://doi.org/10.3390/jcm9061981.
[48] Ahn SH, Yeo AU, Kim KH, Kim C, Goh Y, Cho S, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. Radiat Oncol Lond Engl 2019;14:213. https://doi.org/10.1186/s13014-019-1392-z.
[49] Chen C, Zheng A, Ou X, Wang J, Ma X. Comparison of radiomics-based machine-learning classifiers in diagnosis of glioblastoma from primary central nervous system lymphoma. FrontOncologia 2020:10. https://doi.org/10.3389/fonc.2020.01151.
[50] Zou F-W, Tang Y-F, Liu C-Y, Ma J-A, Hu C-H. Concordance study between IBM Watson for oncology and real clinical practice for cervical cancer patients in China: a retrospective analysis. Front Genet 2020:11. https://doi.org/10.3389/fgene.2020.00200.
[51] Ladefoged CN, Marner L, Hindsholm A, Law I, Højgaard L, Andersen FL. Deep learning based attenuation correction of PET/MRI in pediatric brain tumor patients: evaluation in a clinical setting. Front Neurosci 2019:13. https://doi.org/10.3389/fnins.2018.01005.
[52] Mergen V, Kobe A, Blüthgen C, Euler A, Flohr T, Frauenfelder T, et al. Deep learning for automatic quantification of lung abnormalities in COVID-19 patients: first experience and correlation with clinical parameters. Eur J Radiol Open 2020:7. https://doi.org/10.1016/j.ejro.2020.100272.
[53] Yang S, Jiang L, Cao Z, Wang L, Cao J, Feng R, Deep learning for detecting corona virus disease, et al. (COVID-19) on high-resolution computed tomography: a pilot study. Ann Transl Med 2019;2020:8. https://doi.org/10.21037/atm.2020.03.132.
[54] Im H, Pathania D, McFarland PJ, Sohani AR, Degani I, Allen M, et al. Design and clinical validation of a point-of-care device for the diagnosis of lymphoma via contrast-enhanced microholography and machine learning. Nat Biomed Eng 2018;2:666–74. https://doi.org/10.1038/s41551-018-0265-3.
[55] Ohta Y, Yunaga H, Kitao S, Fukuda T, Ogawa T. Detection and classification of myocardial delayed enhancement patterns on mr images with deep neural networks: a feasibility study. RadiolArtif Intell 2019:1. https://doi.org/10.1148/ryai.2019180061.
[56] Birkenbihl C, Emon MA, Vrooman H, Westwood S, Lovestone S, Hofmann-Apitius M, et al. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia - lessons for translation into clinical practice. EPMA J 2020;11:367–76. https://doi.org/10.1007/s13167-020-00216-z.
[57] Liu Z, Li L, Li T, Luo D, Wang X, Luo D. Does a deep learning-based computer-assisted diagnosis system outperform conventional double Reading by radiologists in distinguishing benign and malignant lung Nodules? Front. Oncologia 2020:10. https://doi.org/10.3389/fonc.2020.545862.
[58] Dembrower K, Wåhlin E, Liu Y, Salim M, Smith K, Lindholm P, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet Digit Health 2020;2:e468–74. https://doi.org/10.1016/S2589-7500(20)30185-0.
[59] Benjamins JW, van Leeuwen K, Hofstra L, Rienstra M, Appelman Y, Nijhof W, et al. Enhancing cardiovascular artificial intelligence (AI) research in the Netherlands: CVON-AI consortium. Neth Heart J 2019;27:414–25. https://doi.org/10.1007/s12471-019-1281-y.
[60] Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M-L, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic

retinopathy in primary care. JAMA Netw Open 2018;1:e182665. https://doi.org/10.1001/jamanetworkopen.2018.2665.
[61] Brunenberg EJL, Steinseifer IK, van den Bosch S, Kaanders JHAM, Brouwer CL, Gooding MJ, et al. External validation of deep learning-based contouring of head and neck organs at risk. Phys Imaging Radiat Oncol 2020;15:8–15. https://doi.org/10.1016/j.phro.2020.06.006.
[62] Böttcher B, Beller E, Busse A, Cantré D, Yücel S, Öner A, et al. Fully automated quantification of left ventricular volumes and function in cardiac MRI: clinical evaluation of a deep learning-based algorithm. Int J Cardiovasc Imaging 2020;36:2239–47. https://doi.org/10.1007/s10554-020-01935-0.
[63] Krause J, Gulshan V, Rahimy E, Karth P, Widner K, Corrado GS, et al. Grader variability and the importance of reference standards for evaluating machine learning models for diabetic retinopathy. Ophthalmology 2018;125:1264–72. https://doi.org/10.1016/j.ophtha.2018.01.034.
[64] Mayo RC, Leung JWT. Impact of artificial intelligence on women's imaging: cost-benefit analysis. Am J Roentgenol 2019;212:1172–3. https://doi.org/10.2214/AJR.18.20419.
[65] Ihlen EAF, Støen R, Boswell L, de Regnier R-A, Fjørtoft T, Gaebler-Spira D, et al. Machine learning of infant spontaneous movements for the early prediction of cerebral palsy: a multi-site cohort study. J Clin Med 2020:9. https://doi.org/10.3390/jcm9010005.
[66] Mehralivand S, Harmon SA, Shih JH, Smith CP, Lay N, Argun B, et al. Multicenter multireader evaluation of an artificial intelligence-based attention mapping system for the detection of prostate cancer with multiparametric MRI. AJR Am J Roentgenol 2020;215:903–12. https://doi.org/10.2214/AJR.19.22573.
[67] Winkel DJ, Breit H-C, Shi B, Boll DT, Seifert H-H, Wetterauer C. Predicting clinically significant prostate cancer from quantitative image features including compressed sensing radial MRI of prostate perfusion using machine learning: comparison with PI-RADS v2 assessment scores. Quant Imaging Med Surg 2020;10:808–23. https://doi.org/10.21037/QIMS.2020.03.08.
[68] Pennig L, Hoyer UCI, Goertz L, Shahzad R, Persigehl T, Thiele F, et al. Primary central nervous system lymphoma: clinical evaluation of automated segmentation on multiparametric MRI using deep learning. J Magn Reson Imaging 2021;53:259–68. https://doi.org/10.1002/jmri.27288.
[69] Rudie JD, Rauschecker AM, Xie L, Wang J, Duong MT, Botzolakis EJ, et al. Subspecialty-level deep gray matter differential diagnoses with deep learning and Bayesian networks on clinical brain mri: a pilot study. Radiol Artif Intell 2020;2:1–13. https://doi.org/10.1148/ryai.2020190146.
[70] Potash E, Ghani R, Walsh J, Jorgensen E, Lohff C, Prachand N, et al. Validation of a machine learning model to predict childhood Lead poisoning. JAMA Netw Open 2020;3:e2012734. https://doi.org/10.1001/jamanetworkopen.2020.12734.
[71] Balidis M, Papadopoulou I, Malandris D, Zachariadis Z, Sakellaris D, Asteriadis S, et al. Validation of neural network predictions for the outcome of refractive surgery for myopia. Med Hypothesis Discov Innov Ophthalmol 2020;9:172–8.
[72] Stuckey TD, Gammon RS, Goswami R, Depta JP, Steuter JA, Meine 3rd Frederick J. Cardiac Phase Space Tomography: A novel method of assessing coronary artery disease utilizing machine learning. PLoS One 2018;13:e0198603. https://doi.org/10.1371/journal.pone.0198603.
[73] Sun C, Zhang Y, Chang Q, Liu T, Zhang S, Wang X, et al. Evaluation of a deep learning-based computer-aided diagnosis system for distinguishing benign from malignant thyroid nodules in ultrasound images. Med Phys Lancet 2020;47:3952–60. https://doi.org/10.1002/mp.14301.
[74] Liu Yuwei, Liu Changqing, Gao Min, Wang Yan, Bai Yangjing, Xu Ruihua, et al. Evaluation of a wearable wireless device with artificial intelligence, iThermonitor WT705, for continuous temperature monitoring for patients in surgical wards: a prospective comparative study. BMJ Open 2020;10:e039474. https://doi.org/10.1136/bmjopen-2020-039474.
[75] Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenogrouping in heart failure to identify responders to cardiac resynchronization therapy. Eur J Heart Fail 2019;21:74–85. https://doi.org/10.1002/ejhf.1333.
[76] Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med 2019;25:70–4. https://doi.org/10.1038/s41591-018-0240-2.
[77] Cesaretti M, Brustia R, Goumard C, Cauchy F, Poté N, Dondero F, et al. Use of artificial intelligence as an innovative method for liver graft macrosteatosis assessment. Liver Transpl 2020;26:1224–32. https://doi.org/10.1002/lt.25801.
[78] Shah P, Mishra DK, Shanmugam MP, Doshi B, Jayaraj H, Ramanjulu R. Validation of deep convolutional neural network-based algorithm for detection of diabetic retinopathy - artificial intelligence versus clinician for screening. Indian J Ophthalmol 2020;68:398–405. https://doi.org/10.4103/ijo.IJO_966_19.
[79] Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018;15:e1002683. https://doi.org/10.1371/journal.pmed.1002683.
[80] Wei Y, Wang W, Cheng M, Hong Z, Gu L, Niu J, et al. Clinical evaluation of a real-time optoelectronic device in cervical cancer screening. Eur J Obstet Gynecol Reprod Biol 2021;266:182–6. https://doi.org/10.1016/j.ejogrb.2021.09.027.
[81] Martins Jarnalo CO, Linsen PVM, Blazís SP, van der Valk PHM, Dickerscheid DBM. Clinical evaluation of a deep-learning-based computer-aided detection system for the detection of pulmonary nodules in a large teaching hospital. Clin Radiol 2021;76:838–45. https://doi.org/10.1016/j.crad.2021.07.012.
[82] Yeh E, Wong E, Tsai C-W, Gu W, Chen P-L, Leung L, et al. Detection of obstructive sleep apnea using belun sleep platform wearable with neural network-based algorithm and its combined use with STOP-bang questionnaire. PLoS One 2021;16:e0258040. https://doi.org/10.1371/journal.pone.0258040.
[83] Jefferies JL, Spencer AK, Lau HA, Nelson MW, Giuliano JD, Zabinski JW, et al. A new approach to identifying patients with elevated risk for fabry disease using a machine learning algorithm. Orphanet J Rare Dis 2021;16:518. https://doi.org/10.1186/s13023-021-02150-3.
[84] Wu Y, Kang K, Han C, Wang S, Chen Q, Chen Y, et al. A blind randomized validated convolutional neural network for auto-segmentation of clinical target volume in rectal cancer patients receiving neoadjuvant radiotherapy. Cancer Med 2022;11:166–75. https://doi.org/10.1002/cam4.4441.
[85] Medina R, Bouhaben J, de Ramón I, Cuesta P, Antón-Toro L, Pacios J, et al. Electrophysiological brain changes associated with cognitive improvement in a pediatric attention deficit hyperactivity disorder digital artificial intelligence-driven intervention: randomized controlled trial. J Med Internet Res 2021;23:e25466. https://doi.org/10.2196/25466.
[86] Schwendicke F, Rossi JG, Göstemeyer G, Elhennawy K, Cantu AG, Gaudin R, et al. Cost-effectiveness of artificial intelligence for proximal caries detection. J Dent Res 2021;100:369–76. https://doi.org/10.1177/0022034520972335.
[87] Tseng AS, Thao V, Borah BJ, Attia IZ, Medina Inojosa J, Kapa S, et al. Cost effectiveness of an electrocardiographic deep learning algorithm to detect asymptomatic left ventricular dysfunction. Mayo Clin Proc 2021;96:1835–44. https://doi.org/10.1016/j.mayocp.2020.11.032.
[88] Hong L, Cheng X, Zheng D. Application of artificial intelligence in emergency nursing of patients with chronic obstructive pulmonary disease. Contrast Media Mol Imaging 2021;2021:6423398. https://doi.org/10.1155/2021/6423398.
[89] Grzybowski A, Brona P. Analysis and comparison of two artificial intelligence diabetic retinopathy screening algorithms in a pilot study: IDx-DR and retinalyze. J Clin Med 2021;10:2352. https://doi.org/10.3390/jcm10112352.
[90] Castillo TJM, Starmans MPA, Arif M, Niessen WJ, Klein S, Bangma CH, et al. A multi-center, multi-vendor study to evaluate the generalizability of a radiomics model for classifying prostate cancer: high grade vs. Low Grade. Diagn Basel Switz 2021;11:369. https://doi.org/10.3390/diagnostics11020369.
[91] McLouth J, Elstrott S, Chaibi Y, Quenet S, Chang PD, Chow DS, et al. Validation of a deep learning tool in the detection of intracranial hemorrhage and large vessel occlusion. Front Neurol 2021;12:656112. https://doi.org/10.3389/fneur.2021.656112.
[92] Joo B, Choi HS, Ahn SS, Cha J, Won SY, Sohn B, et al. A deep learning model with high standalone performance for diagnosis of unruptured intracranial aneurysm. Yonsei Med J 2021;62:1052–61. https://doi.org/10.3349/ymj.2021.62.11.1052.
[93] Connolly P, Stapleton S, Mosoyan G, Fligelman I, Tonar Y-C, Fleming F, et al. Analytical validation of a multi-biomarker algorithmic test for prediction of progressive kidney function decline in patients with early-stage kidney disease. Clin Proteomics 2021;18:26. https://doi.org/10.1186/s12014-021-09332-y.
[94] Haverinen J, Keränen N, Falkenbach P, Maijala A, Kolehmainen T, Reponen J. Digi-HTA: health technology assessment framework for digital healthcare services. Finn J EHealth EWelfare 2019;11:326–41. https://doi.org/10.23996/fjhw.82538.
[95] Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat Med 2021;27:582–4. https://doi.org/10.1038/s41591-021-01312-x.
[96] Cutillo CM, Sharma KR, Foschini L, Kundu S, Mackintosh M, Mandl KD. Machine intelligence in healthcare—perspectives on trustworthiness, explainability, usability, and transparency. Npj Digit Med 2020;3:1–5. https://doi.org/10.1038/s41746-020-0254-2.
[97] Navarro CLA, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. BMJ 2021;375:n2281. https://doi.org/10.1136/bmj.n2281.
[98] Tariq A, Purkayastha S, Padmanaban GP, Krupinski E, Trivedi H, Banerjee I, et al. Current clinical applications of artificial intelligence in radiology and their best supporting evidence. J Am Coll Radiol 2020;17:1371–81. https://doi.org/10.1016/j.jacr.2020.08.018.
[99] Retson TA, Eghtedari M. Computer-aided Detection/Diagnosis in breast imaging: a focus on the evolving FDA regulations for using software as a medical device. Curr Radiol Rep 2020:8. https://doi.org/10.1007/s40134-020-00350-6.
[100] Martelli N, Eskenazy D, Déan C, Pineau J, Prognon P, Chatellier G, et al. New european regulation for medical devices: what is Changing? Cardiovasc Intervent Radiol 2019;42:1272–8. https://doi.org/10.1007/s00270-019-02247-0.
[101] Stern AD, Price WN. Regulatory oversight, causal inference, and safe and effective health care machine learning. Biostat Oxf Engl 2020;21:363–7. https://doi.org/10.1093/biostatistics/kxz044.
[102] Markus AF, Kors JA, Rijnbeek PR. 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 2021;113:103655. https://doi.org/10.1016/j.jbi.2020.103655.
[103] Park SH, Kressel HY. Connecting technological innovation in artificial intelligence to real-world medical practice through rigorous clinical validation: what peer-reviewed medical journals could do. J Korean Med Sci 2018;33:e152. https://doi.org/10.3346/jkms.2018.33.e152.
[104] Magrabi F, Ammenwerth E, McNair JB, De Keizer NF, Hyppönen H, Nykänen P, et al. Artificial intelligence in clinical decision support: challenges for evaluating AI and practical implications. Yearb Med Inform 2019;28:128–34. https://doi.org/10.1055/s-0039-1677903.
[105] Ethics guidelines for trustworthy AI | Shaping Europe's digital future n.d. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed October 20, 2021).

[106] Hanif AM, Beqiri S, Keane PA, Campbell JP. Applications of interpretability in deep learning models for ophthalmology. Curr Opin Ophthalmol 2021;32:452–8. https://doi.org/10.1097/ICU.0000000000000780.
[107] Goldsack JC, Zanetti CA. Defining and developing the workforce needed for success in the digital era of medicine. Digit Biomark 2020;4:136–42. https://doi.org/10.1159/000512382.
[108] He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30–6. https://doi.org/10.1038/s41591-018-0307-0.
[109] Chiang S, Picard RW, Chiong W, Moss R, Worrell GA, Rao VR, et al. Guidelines for conducting ethical artificial intelligence research in neurology: a systematic approach for clinicians and researchers. Neurology 2021;97:632–40. https://doi.org/10.1212/WNL.0000000000012570.
[110] Alami H, Lehoux P, Denis J-L, Motulsky A, Petitgand C, Savoldelli M, et al. Organizational readiness for artificial intelligence in health care: insights for decision-making and practice. J Health Organ Manag 2020;35:106–14. https://doi.org/10.1108/JHOM-03-2020-0074.
[111] Yirmibesoglu Erkal E, Akpınar A, Erkal HŞ. Ethical evaluation of artificial intelligence applications in radiotherapy using the four topics approach. Artif Intell Med 2021;115:102055. https://doi.org/10.1016/j.artmed.2021.102055.
[112] Martinho A, Kroesen M, Chorus C. A healthy debate: exploring the views of medical doctors on the ethics of artificial intelligence. Artif Intell Med 2021;121:102190. https://doi.org/10.1016/j.artmed.2021.102190.
[113] Jayakumar S, Sounderajah V, Normahani P, Harling L, Markar SR, Ashrafian H, et al. Quality assessment standards in artificial intelligence diagnostic accuracy systematic reviews: a meta-research study. Npj Digit Med 2022;5:1–13. https://doi.org/10.1038/s41746-021-00544-y.
[114] Saginur M, Fergusson D, Zhang T, Yeates K, Ramsay T, Wells G, et al. Journal impact factor, trial effect size, and methodological quality appear scantly related: a systematic review and meta-analysis. Syst Rev 2020;9:53. https://doi.org/10.1186/s13643-020-01305-w.
[115] Farah L, Murris JM, Borget I, Guilloux A, Martelli NM, Katsahian SIM. Assessment of performance, interpretability, and explainability in artificial intelligence–based health technologies: what healthcare stakeholders need to know. Mayo Clinic Proceedings: Digital Health 2023;1:120–38. https://doi.org/10.1016/j.mcpdig.2023.02.004.