RESEARCH

Open Access

Landscape and challenges in economic evaluations of artificial intelligence in healthcare: a systematic review of methodology

Nanna Kastrup[1*], Annette W. Holst-Kristensen[1] and Jan B. Valentin[1]

Abstract

Background The potential for artificial intelligence (AI) to transform healthcare cannot be ignored, and the development of AI technologies has increased significantly over the past decade. Furthermore, healthcare systems are under tremendous pressure, and efficient allocation of scarce healthcare resources is vital to ensure value for money. Health economic evaluations (HEEs) can be used to obtain information about cost-effectiveness. The literature acknowledges that the conduct of such evaluations differs between medical technologies (MedTechs) and pharmaceuticals, and poor quality evaluations can provide misleading results. This systematic review seeks to map the evidence on the general methodological quality of HEEs for AI technologies to identify potential areas which can be subject to quality improvements. We used the 35-item checklist by Drummond and Jefferson and four additional checklist domains proposed by Terricone et al. to assess the methodological quality of full HEEs of interventions that include AI.

Results We identified 29 studies for analysis. The included studies had higher completion scores for items related to study design than for items related to data collection and analysis and interpretation of results. However, none of the studies addressed MedTech-specific items.

Conclusions There was a concerningly low number of full HEEs relative to the number of AI publications, however the trend is that the number of studies per year is increasing. Mapping the evidence of the methodological quality of HEEs of AI shows a need to improve the quality in particular the use of proxy measures as outcome, reporting, and interpretation of the ICER.

Keywords Artificial intelligence, Health economic evaluation, Cost-effectiveness, Cost-utility analysis, Cost-effectiveness analysis, Ssystematic review

Introduction

The rapid adoption of medical technologies (MedTechs), including artificial intelligence (AI) solutions, in healthcare is consuming valuable resources. However, knowledge about their cost-effectiveness is limited [1]. AI in healthcare is often used as decision support, and its effect

*Correspondence:
Nanna Kastrup
nkh@dcm.aau.dk
[1] Department of Clinical Medicine, Danish Center for Health Services Research, Aalborg University, Aalborg, Denmark

on clinical outcomes can be mediated by clinicians'. Consequently, the academic discipline of validating decision support systems as standard prediction models may be insufficient [2, 3]. Therefore, clinical intervention studies on effect and cost-effectiveness are warranted. The academic discipline of health economic evaluations (HEE) focuses on how scarce healthcare resources can be efficiently allocated to maximise health [4]. Such evaluations aim to inform decision-makers about cost-effective courses of action when comparing two or more interventions [4]. The methodological framework used for estimating the cost-effectiveness of pharmaceuticals is well established and adopted by the industry, regulatory bodies, and in research. The results from such analyses are often used in prioritising and health technology assessment processes [5].

New technologies are developed and constantly introducing new treatment paradigms. Studies reflect on how cost-effectiveness can be impacted by temporal learning curve differences, continuous product modifications, dynamic pricing, and how adoption often requires substantial investments and re-organisation, which are distinctive characteristics frequently related to MedTechs, including AI [6, 7]. In general, AI performs different compared to devices and pharmaceuticals since they are purely data driven non-invasive technologies.

AI shows the potential to promote health in diverse disciplines, such as the development of genetic medicine, diagnostics, and predictive analytics. Its development trend is strong, with the number of new AI technologies exceeding the number of new pharmaceuticals in 2020 [8, 9]. Current guidelines recommends that AI applications are evaluated almost similar to pharmaceuticals [2] . However, given the nature of AI, the evaluation of such applications are traditionally reduced to assessing predictive performance [10, 11]. Unsurprisingly, a systematic review reported that the evidence for AI's cost-effectiveness in healthcare is limited and based on relatively few evaluations ($n = 20$) compared to the total number of studies describing AI in healthcare, which was 120,000 for 2019 alone [1, 12]. Furthermore, 50% of HEEs did not report details on the analytic method, model assumptions, or characterising uncertainty, and the reported outcomes predominantly focused on costs [1]. Given the differences between evaluating pharmaceuticals and MedTech's and the rapid development of AI solutions, conducting high-quality economic evaluations is critical [13]. The worst-case scenario is that the primary purpose of conducting an HEE is undermined and misguided. This systematic review seeks to map the evidence of the general methodological quality of HEEs for AI technologies. This is done by screening the scientific literature, which evaluates cost-effectiveness of AI interventions and assessing the methodological quality using the 35-item checklist by Drummond and Jefferson and four additional checklist domains proposed by Terricone et al. [14, 15]. The result from the study is sought to inform decision-makers about the current landscape of the methodological quality of recent published studies and raise the awareness of domains which can be subject to quality improvements.

Methods

Design

We systematically reviewed the methodology for full HEEs of clinical interventions for AI implementations in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses-guidelines (PRISMA) [16].

Search strategy

A systematic literature search was performed in the Embase, PubMed, Web of Science, and Scopus databases based on the recommendations for systematic reviews of economic evaluations by Ghislaine et al. [17]. The search included peer-reviewed, full-text studies published within the last five years, between 1 January 2017 and 6 March 2022, and updated until 1 September 2023. Relevant studies with the following health economic terms in their abstract or title were identified: 'economic evaluation' OR 'cost-effectiveness' OR 'cost-utility' OR 'cost-benefit'. Search terms for AI were: 'artificial intelligence' OR 'machine learning' OR 'deep learning' OR 'decision support systems'. Where possible, Medical Subject Headings (MeSH) terms were used. A pilot test was conducted to assess whether the databases' search filters could further limit the search. Since they excluded already-known studies, a broad search strategy was used. The full search query for each database, number of hits, and access date are provided in the Supplementary Material. After conducting the final search, duplicates were automatically removed first in Endnote and then by manual comparison of title, journal, author, and publication year in rayan.ai.

Inclusion and exclusion criterions

Only clinical intervention studies were included. These studies were eligible for inclusion if they included at least one comparator based on AI and conducted at least one full HEE. A full HEE was defined by Drummond et al. as *"a comparative analysis of alternative courses of action in terms of both their costs and consequences"* and comprises cost-effectiveness analyses (CEAs), cost-utility analyses (CUAs), and cost-benefit analyses (CBAs) [4]. The measure of effect differs between the three types of analyses, however, the methodological approach for assessment is the same, and therefore appropriate to apply across disease areas. Early-stage HEEs were excluded to avoid misleading results since they require adjustments before final cost-effectiveness estimation. The articles' language was restricted to English and Danish.

Selection of sources of evidence

Two of the authors (NK and AWHK) conducted the screening of abstracts and assessment of study eligibility. The two authors screened independently and blinded from one another using rayan.ai [18]. Duplicate records were removed automatically prior to screening using Endnote and rayan.ai. Abstracts were accepted for full text assessment if both authors independently agreed to include and removed from further assessment if both authors independently agreed to exclude. Situations where the two authors disagreed or one or both authors were in doubt was handled by discussion until a consensus was reached. If a consensus could not be reached the paper was included for full-text assessment. The final inclusion of studies was determined after in-dept reading of full texts against inclusion criteria. Once more, any disagreements were discussed until a consensus was reached.

Data extraction

Extraction of data was done independently by two authors (NK and AWHK). A data collection spreadsheet was developed, encompassing an extensive process of testing to ascertain the accuracy of data extraction. Through a series of iterative stages, this testing formed the final standardized data collection template. Furthermore, comprehensive discussions were conducted, leading to a consensus among reviewers regarding the terminology and definitions to be employed. Additionally, each individual item on the checklist from Drummond et al. was subjected to detailed inspection, along with the supplementary document providing elaboration for each checklist item. The following study characteristics are reported: author, year, country, objective, clinical domain, and self-reported type of AI intervention term. Furthermore, we extracted the studies' health economic characteristics: settings, study perspective(s), evaluation type, decision-analytic model (DAM) or alongside clinical trial (ACL), primary outcome measure, HEE type, incremental cost-effectiveness ratio (ICER), cost-effective alternative, and probabilistic sensitivity analysis (PSA). NK and AWHK independently extracted all data.

Quality assessment

Quality assessment was conducted following carefully selected checklists. The checklists were chosen after reviewing 13 identified checklists [19, 20]. The criteria for the checklist included applicability to full HEEs based on ACL and DAM and to assessment of both CEA, CUA and CBA. The Drummond 35-item checklist matched these criteria [14]. The Cochrane Handbook for Systematic Reviews of Interventions also recommends this checklist for critically appraising methodological quality [21]. Each item was interpreted as described by Drummond et al. [14]. The checklist comprises three categories: study design (items 1–7), data collection (items 8–21), and analysis and interpretation of results (items 22–35). Items are completed using a 'yes', 'no', 'cannot tell', or 'not applicable' scale [14]. We also identified and added four supplementary items suggested as relevant for MedTechs [15]. The results for the supplementary items were reported similarly to Drummond's checklist, with the answers recorded as 'formal', 'substantial', or 'not stated' [15].

Results

A total of 13.319 records were identified. After duplicate removal, 7459 unique records were retained for further review. Three systematic reviews were identified for snowballing, which did not lead to the inclusion of additional studies [1, 13, 22]. In all, 7420 studies did not meet the inclusion criteria for full text screening. Forthy studies were subjected to full-text screening, of which thirteen were subsequently excluded since they did not meet the criteria. The excluded studies were cost-minimisation analyses, were claiming to be CEAs but only included costs, or were unavailable in English or Danish. Details are provided in Fig. 1. In all, 27 studies are identified and the included study by Rossi et al. encompassed three unique HEEs, which we assessed independently. Therefore, 17 HEEs were assessed from the 15 records identified [23].

Studies' general characteristics

A general overview of the study design is provided in Table 1. We identified 8 (28%) studies published in 2023, 13 studies (45%) published in 2022, five (17%) in 2021, only one (3%) in 2019, 2017, and 2020. No studies were identified in 2018. Ten HEEs (34%) were conducted in the US, three (10%) in Germany and China, two (7%) in the UK, and 11 (38%) in other countries. Eleven (38%) HEEs reported the generic AI term to describe the AI technology, six studies (21%) used convolutional neural network (CNN) or machine learning (ML), and five studies (17%) uses deep learning (DL), and one study (3%) reports using an artificial neural network.

Studies' health economic characteristics

A general overview of health economic features is provided in Table 2. Remarkably, 13 studies (45%) did not

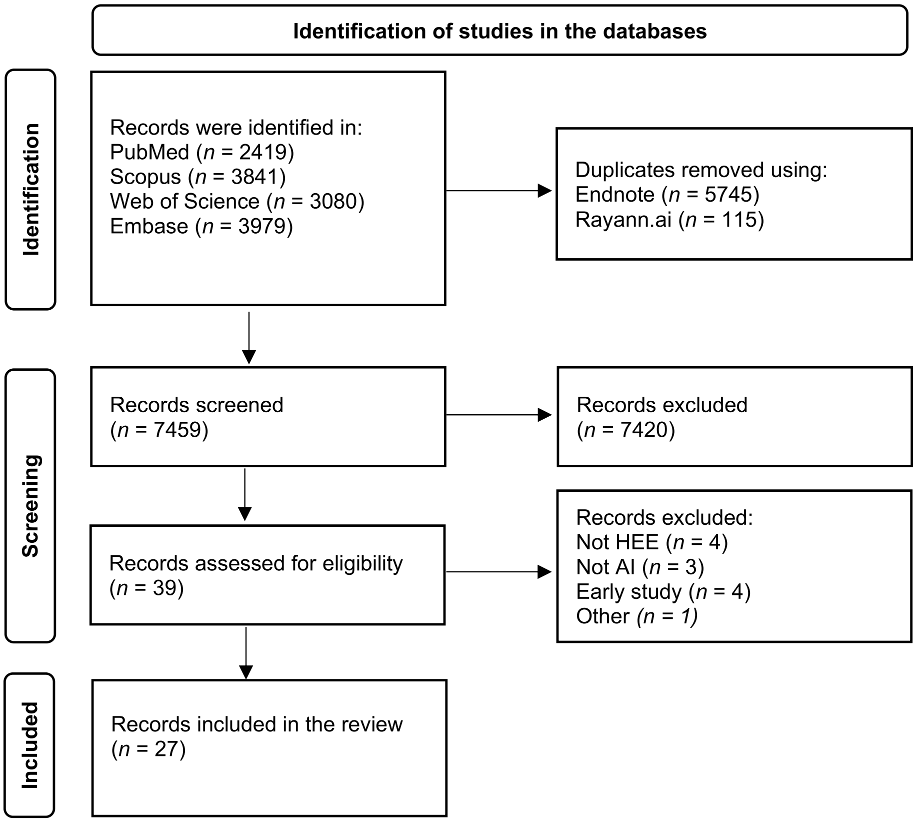
Identification of studies in the databases
Identification
Records were identified in:
PubMed (*n* = 2419)
Scopus (*n* = 3841)
Web of Science (*n* = 3080)
Embase (*n* = 3979)
Duplicates removed using:
Endnote (*n* = 5745)
Rayann.ai (*n* = 115)
Screening
Records screened (*n* = 7459)
Records excluded (*n* = 7420)
Records assessed for eligibility (*n* = 39)
Records excluded:
Not HEE (*n* = 4)
Not AI (*n* = 3)
Early study (*n* = 4)
Other (*n = 1*)
Included
Records included in the review (*n* = 27)

Fig. 1 A PRISMA chart of the identification, screening, and inclusion of full-text records

report information on study settings. Regarding study perspectives, healthcare (*n* = 13, 45%) and payer perspectives (*n* = 7, 24%) predominated, where three studies (10%) use a combination of perspectives, and one (3%) uses a family perspective. For HEE types the CUA (*n* = 19, 65%) is predominant compared with CEA (*n* = 9, 31%), notably from the year 2023 and forth only CUA was used. The Markov model (*n* = 20, 69%) is the preferred method to analyse of cost-effectiveness, and six HEEs (21%) used a combination of decision trees and Markov models. Four HEEs (14%) used ACLs, which were all CEAs. In all, 18 HEEs (62%) reported the base-case ICER, but notably, 11 studies (29%) did not report a base-case ICER clearly, and seven studies (24%) reported more than one base-case ICER. Two HEEs (7%) reported the control cost-effective, and 20 (69%) deemed the intervention cost-effective. A complete overview of the details concerning HEE characteristics is provided in Table 3. Additional variables for the country, time horizon, currency and price index year, number of alternatives, discounting rates for cost and effect, types of sensitivity analysis, reporting of methods for comparison of more than two alternatives, ICER plot in the incremental cost-effectiveness (ICE) plane, willingness-to-accept, and willingness-to-pay are presented in Table S1.

Quality assessment

The checklist items used for quality assessment are shown in Table 3 with corresponding response frequencies. Regarding items on study design (1–7), most HEEs (*n* = 26, 90%) provided well-described research questions (item 1) [23–46]. However, while three HEEs (10%) did not clearly state which comparator they applied in their research question, the alternatives were described elsewhere (item 5) [47–49]. Moreover, eight HEEs (28%) stated the economic importance of actual cost quantifications (item 2) [26–28, 32, 43, 45, 48, 49], while the others only used generic statements similar to 'costly' [23–25, 29–31, 33–42, 44, 46, 47]. Viewpoint was stated in all but three HEE (item 3) [34, 38, 41]. All studies report a clear description of the alternatives being compared as well as the form of economic evaluation (item 6), however, none of the studies give a justification on the choice of HEE among CEA, CUA or CBA, nor did any of the studies give a rational for the choice of effectiveness or utility measure where appropriate (item 7).

Table 1 General characteristics of the included HEEs ($n = 29$)

Author	Year	Country	Objective	Clinical domain	AI term
Mervin et al.	2017	Australia	To examine the within-trial costs and cost-effectiveness of using PARO, compared with a plush toy and usual care, for reducing agitation and medication use in people with dementia in long-term care.	Dementia	AI
Padula et al.	2019	US	To analyse the cost-utility of performing repeated risk assessment for pressure-injury prevention in all patients or high-risk groups.	Pressure injury	ML
Hill et al.	2020	UK	To assess the cost-effectiveness of targeted screening, informed by a machine learning risk prediction algorithm, to identify patients with atrial fibrillation.	Atrial fibrillation	ML
Nsengiyumva et al.	2021	Pakistan	To evaluate the cost-effectiveness of triage strategies with artificial intelligence-based chest X-ray analysis for patients with symptoms suggestive of pulmonary tuberculosis.	Tuberculosis	DL-based AI
Schwendicke et al.	2021	Germany	To assess the cost-effectiveness of using artificial intelligence for proximal caries detection on bitewing radiographs.	Caries detection	CNN
Tseng et al.	2021	US	To evaluate the cost-effectiveness of an artificial intelligence-electrocardiogram algorithm under various clinical and cost scenarios when used for universal screening at age 65.	Asymptomatic left ventricular dysfunction	AI
Turino et al.	2021	Spain	To assess the effectiveness and cost-effectiveness of an intelligent monitoring system for improving continuous positive airway pressure compliance.	Obstructive sleep apnoea	ML
Mallow et al.	2021	US	To conduct a cost-utility analysis of a novel genetic diagnostic test (opioid use disorder test) for assessing the risk of developing opioid use disorder in elective orthopaedic surgery patients.	Opioid use disorder	ML
Delgadillo et al.	2022	UK	To compare the clinical and cost effectiveness of two treatment selection strategies: stepped and stratified care.	Psychological treatment	ML
Fuller et al.	2022	US	A cost-effectiveness analysis of an automated retinal image analysis system-based diabetic retinopathy screening in a primary care medical clinic serving a low-income patient population.	Diabetic retinopathy	AI
Rossi et al	2022	US	To assess the cost-effectiveness of artificial intelligence for supporting clinicians in detecting and grading diseases in dermatology.	Dermatology	CNN
Rossi et al.	2022	Germany	To assess the cost-effectiveness of artificial intelligence for supporting clinicians in detecting and grading diseases in dentistry.	Dentistry	CNN
Rossi et al.	2022	Brazil	To assess the cost-effectiveness of artificial intelligence for supporting clinicians in detecting and grading diseases in ophthalmology.	Ophthalmology	CNN
Huang et al.	2022	China	To assess the cost-effectiveness of artificial intelligence screening for diabetic retinopathy compared to conventional screening strategies.	Diabetic retinopathy	AI
Morrison et al.	2022	US	To evaluate the relative cost-effectiveness of autonomous and assistive artificial intelligence-based retinopathy of prematurity screening compared with telemedicine and ophthalmoscopic screening over a range of estimated probabilities, costs, and outcomes.	Retinopathy	AI
Schwendicke et al.	2022	Germany	To assess the cost-effectiveness of artificial intelligence-supported detection of proximal caries in a randomised controlled clustered cross-over superiority trial.	Caries detection	CNN

Table 1 (continued)

Author	Year	Country	Objective	Clinical domain	AI term
Wolf et al.	2022	US	To assess the cost-effectiveness of detecting and treating diabetic retinopathy and its sequelae among children with type 1 diabetes and type 2 diabetes using artificial intelligence diabetic retinopathy screening vs standard screening by an eye care professional.	Diabetic retinopathy	AI
Hill et al.	2022	England	To determine the effectiveness of a screening strategy that included a machine learning risk prediction algorithm in conjunction with diagnostic testing for identification of undiagnosed atrial fibrillation.	Atrial fibrillation	ML
Mital et al	2022	US	To examine the cost-effectiveness of using Artificial intelligence or polygenic risk score to guide mammography screening for breast cancer compared with screening based exclusively on family history, annual screening for all women and no screening, among white women.	Breast cancer	DL
Skarping et al	2022	Sweden	To evaluate the utility of the non-invasive lymph node staging model in reducing the proportion of cN0 patients with low predicted risk undergoing sentinel lymph node biopsy.	Brest cancer screening	ANN
Ziegelmayer et al.	2022	US	To evaluate the cost-effectiveness of using an artificial intelligence algorithm for initial screening scan for lung cancer.	Lung cancer screening	DL
Barkun et al.	2023	Canada	To estimate incremental cost-effectiveness ratio comparing computer-assisted diagnosis to conventional colonoscopy polyp detection amongst patients with a positive faecal immunochemical test.	Colonoscopy	AI
Chawla et al.	2023	US	To compare the automated versus manual screening and management pathway for diabetic patients with unknown retinopathy status.	Diabetic retinopathy	DL
Hassan et al.	2023	Italy	To analyse the cost-effectiveness of computer-assisted diagnosis colonoscopy in the detection of adenomas and colorectal cancer in Italy.	Colonoscopy	AI
Lin et al.	2023	China	The objective of our community-based telemedicine screening for diabetic retinopathy was to examine whether the artificial intelligence model can be more cost-effective than manual grading in low- and middle-income countries.	Diabetic retinopathy	CNN
Pickhardt et al.	2023	CT	To assess the cost-effectiveness and clinical efficacy of artificial intelligence-assisted abdominal computer tomography-based opportunistic screening for atherosclerotic cardiovascular disease, osteoporosis, and sarcopenia using artificial intelligence body composition algorithms.	Atherosclerotic cardiovascular disease and osteoporosis	AI
Shen et al.	2023	China	We aimed to evaluate the cost-effectiveness of artificial intelligence-assisted liquid-based cytology testing, compared with the manual liquid-based cytology and human papilloma virus-DNA testing, for primary cervical cancer screening in China.	Cervical cancer screening	AI
Srisubat et al.	2023	Thailand	This study was conducted to analyse the cost-utility of deep learning compared to trained non-physician human graders to screen diabetic retinopathy over a lifetime horizon of patients	Diabetic retinopathy screening	DL
Yonazu et al.	2023	Japan	To evaluate the cost-effectiveness of a computer-assisted diagnosis system designed to support clinicians in differentiating early gastric cancers	Gastric cancer	AI

Key: *AI* Artificial intelligence, *ANN* Artificial neural network, *ML* Machine learning, *CNN* Conventional neural network, *DL* Deep learning

Table 2 Health economic characteristics of the included HEEs (*n* =29)

Author	Settings	Perspective	HEE type	DAM type or ACL	Primary outcome	ICER	Cost-effective alternative
Delgadillo et al.	Psychological therapy services	Health service	CEA	ACL	PHQ-9 measure	CT	Intervention
Fuller et al.	Primary care	Payer	CUA	Markov	QALY	$258,721.81/QALY	Intervention
Rossi et al.	CT	Payer	CUA	Markov	QALY	–$27,580/QALY	CT
Rossi et al.	CT	Payer	CEA	Markov	Tooth retention time	–€15.01/year	CT
Rossi et al.	CT	Payer	CUA	Markov	QALY	R-$91,760/QALY	CT
Hill et al.	Primary care	National health service	CUA	Decision tree and Markov	QALY	Systematic: £4,847/QALY Opportunistic: £5,544/QALY	Intervention
Huang et al.	Health service stations	Health system and societal	CUA	Decision tree and Markov	QALY	Health system: $1,107.63/QALY Societal: $1,0347.12/QALY	Intervention
Mervin et al.	Long-term care facilities	Healthcare	CEA	ACL	CMAI-SF	PARO: $13.01/CMAI-SF Plush toy: $12.85/CMAI-SF	CT
Morrison et al.	CT	Healthcare	CEA	Decision tree	QALY	CT	CT
Nsengiyumva et al.	Health centre	Healthcare	CEA	Decision tree	DALY	CT	CT
Padula et al.	Hospital	Societal and healthcare	CUA	Markov	QALY	Healthcare: $2,142/QALY Societal: $2,000/QALY	Intervention
Schwendicke et al.	Private dental practice	Healthcare	CEA	ACL	Tooth retention time	– €8.9/year	Control
Schwendicke et al.	CT	Public-private-payer healthcare	CEA	Markov	Tooth retention time	– €13.9 /year	Intervention
Tseng et al.	CT	Healthcare	CUA	Decision tree and Markov	QALY	$43,351/QALY	Intervention
Wolf et al.	CT	Family	CEA	Decision tree	True positive proportion	T1DM: $31/TPP T2DM: $95/TPP	Intervention
Turino et al.	Hospital sleep unit	CT	CEA	ACL	Hours/day of CPAP use	CT	Intervention
Mallow et al.	Orthopaedic pain management	Payer and self-insured payer	CUA	Markov	QALY	CT	Intervention
Barkun et al.	Health care	Payer's perspective	CUA	Markov	QALY and LY	CT	Intervention
Chawla et al.	Primary care	Health care payer	CUA	Markov	QALY	CT	Intervention
Hassan et al.	Primary screening	National health service	CUA	Markov	QALY	CT	Intervention
Hill et al.	Primary care	National health service	CUA	Decision tree and Markov	QALY	£3,994/QALY	Intervention
Lin et al.	Community health service center	Societal	CEA and CUA	Markov	Years without blindness per 100,000 people with DM, and QALY	ICER: $2553.39/year ICUR: $15,216.96/QALY	Control
Mital et al.	CT	Health care system	CUA	Decision tree and microsimulation	QALY	$23,755/QALY	Intervention
Pickhardt et al.	CT	CT	CUA	Markov	QALY	CT	Intervention
Shen et al.	Primary screening	Health care provider	CUA	Markov	QALY	$622–24,482/QALY	Intervention

Table 2 (continued)

Author	Settings	Perspective	HEE type	DAM type or ACL	Primary outcome	ICER	Cost-effective alternative
Skarping et al.	CT	Health care	CUA	Decision tree	QALY	CT	Intervention
Srisubat et al.	CT	Health care provider and societal	CUA	Decision tree and Markov	QALY	Provider: $512,955/QALY Societal: CT	CT
Yonazu et al.	CT	Medical payer	CUA	Decision tree and Markov	QALY	$11,0937QALY	Intervention
Ziegelmayer et al.	CT	Health care	CUA	Markov	QALY	CT	Intervention

Key: *HEE* Health economic evaluation, *PHQ-9* Patient health questionnaire, *CT* Cannot tell, *CEA* Cost-effectiveness analysis, *CUA* Cost-utility analysis, *QALY* Quality-adjusted life years, *NHS* National health service, *DALY* Disability-adjusted life year, *TTP* True positive proportion, *ACL* Alongside clinical trial, *T1DM* Type-1 diabetes mellitus, *T2DM* Type-2 diabetes mellitus, *PHQ* Patient health questionnaire, *CPAP* Continuous positive airway pressure, *CMAI-SF* Cohen–Mansfield Agitation Inventory-Short Form, *DAM* Decision-analytic model, *ICER* Incremental cost-effectiveness ratio, *TPP* True positive proportion, *DM* Diabetes mellitus, *ANN* Artificial Neural Network

Regarding items on data collection (8–21), all but one (3 %) [44] HEE clearly stated the primary outcome of the evaluation (item 11) [44] and sources of effectiveness estimation is stated in all studies (item 8). However, only eight HEEs (28%) reported details of the subjects from whom valuation is obtained (item 13) [23, 26, 29, 34, 35, 40, 43]. Nineteen HEEs (66%) [25, 28, 29, 32, 33, 35–48] described the methods for estimation of quantities and unit costs (item 17), but only 12 HEEs (41%) reported recourse use separately from unit costs (item 16) [25, 29, 35, 37–41, 43–45, 47].

Regarding items related to results analysis and interpretation (22–35), only two HEEs (10%) did not report a time horizon (item 22) [24, 33]. Ten HEEs (34%) did not report details of the statistical tests and confidence intervals for stochastic data (item 26) [26, 28, 30, 31, 36, 37, 41, 42, 44, 45]. All but one HEE reported approaches to sensitivity analyses (item 27) and conducted either one-way sensitivity analysis or/and a PSA [28]. Furthermore, four HEEs (14%) did not report an incremental analysis (item 31) [24, 34, 38, 46], which was also unclear in seven HEEs (24%) [29, 36, 37, 41, 44, 47, 49]. Eight HEEs (28%) did not report cost and effects in disaggregated and aggregated forms (item 32) [24, 30, 34, 37, 40, 41, 44, 46]. Three HEEs (10%) do not provide an answer as to which alternative was cost-effective [23, 32, 47–49] and five studies (17%) are not stating it clearly but report statements such as numeric numbers but without conclusion [23, 33, 43] None of the HEEs reported information about the learning curve, incremental innovation, dynamic pricing, or organisational impact. These items included incremental innovation, dynamic pricing, the learning curve, and organisational impact.

Discussion

This systematic review summarised the methodological quality of HEEs that included AI interventions as a comparator. The results are not presented as aggregated study-specific percentage scores since items were considered to be weighted differently according to their impact on methodological quality. Furthermore, the results show that items related to study design and data analysis have higher completion scores than items related to results analysis and interpretation. Surprisingly, none of the studies addressed MedTech-specific items. Another systematic review used the CHEERS checklist to assess the methodological quality of HEEs on AI interventions. They also identified relatively few studies, given the high publication rate of AI studies, and found that most did not report details on analytical methods (n = 14), model assumptions (n = 11), and characterising uncertainty (n = 12) [1]. In this systematic review, all studies provided well-described research questions, the viewpoint of the analyses, and the primary outcome, which is valuable information when understanding the very context of the evaluation. However, when addressing the analytical part of the HEEs, seven studies did not report unit costs and resource use separately. In general, transparency about any input variable is critical to validate the HEE and understand drivers for cost-effectiveness. The time horizon was also missing in three studies. Four studies did not report statistical tests and/or confidence intervals for outcome data, which is concerning for several reasons. First, HEEs are used as input tools for decision-making. Therefore, not stating a time horizon causes fundamental challenges since it prevents the reader from assessing whether it captures major cost and health impact consequences [4]. In addition, not providing information on the statistical uncertainties of stochastic data is problematic since it reflects whether key parameters are candidates for sensitivity analysis in both one-way- or multiway-PSA [4]. However, the three most noteworthy methodological quality drawbacks we found was related to the ICER and effect measures. Firstly, clear reporting of the results of the HHEs in the form of an ICER,

Table 3 Summary of aggregated results for each the items of the Drummond and Jeffersons 35- item checklist

	Yes (%)	No (%)	Can't tell (%)	n/a (%)
Study design				
1. The research question is stated.	26 (90)	0 (0)	3 (10)	
2. The economic importance of the research question is stated.	8 (28)	21 (72)	0 (0)	
3. The viewpoint(s) of the analysis are clearly stated and justified.	26 (90)	3 (10)	0 (0)	
4. The rationale for choosing alternative programmes or interventions compared is stated.	28 (97)	1 (3)	0 (0)	
5. The alternatives being compared are clearly described.	26 (90)	0 (0)	3 (10)	
6. The form of economic evaluation used is stated.	29 (100)	0 (0)	0 (0)	
7. The choice of form of economic evaluation is justified in relation to the questions addressed.	0 (0)	29 (100)	0 (0)	
Data collection				
8. The source(s) of effectiveness estimates used are stated.	29 (100)	0 (0)	0 (0)	0 (0)
9. Details of the design and results of effectiveness study are given (if based on a single study).	4 (14)	0 (0)	0 (0)	25 (86)
10. Details of the methods of synthesis or meta-analysis of estimates are given (if based on a synthesis of a number of effectiveness studies).	23 (79)	0 (0)	0 (0)	6 (21)
11. The primary outcome measure(s) for the economic evaluation are clearly stated.	28 (97)	1 (3)	0 (0)	0 (0)
12. Methods to value benefits are stated.	25 (86)	4 (14)	0 (0)	0 (0)
13. Details of the subjects from whom valuations were obtained were given.	8 (28)	21 (72)	0 (0)	0 (0)
14. Productivity changes (if included) are reported separately.	6 (21)	0 (0)	0 (0)	23 (79)
15. The relevance of productivity changes to the study question is discussed.	3 (10)	3 (10)	1 (3)	22 (76)
16. Quantities of resource use are reported separately from their unit costs.	12 (41)	17 (59)	0 (0)	0 (0)
17. Methods for the estimation of quantities and unit costs are described.	19 (65)	7 (24)	3 (10)	0 (0)
18. Currency and price data are recorded.	22 (76)	0 (0)	7 (24)	0 (0)
19. Details of currency of price adjustments for inflation or currency conversion are given.	22 (76)	7 (24)	0 (0)	0 (0)
20. Details of any model used are given.	29 (100)	0 (0)	0 (0)	0 (0)
21. The choice of model used and the key parameters on which it is based are justified.	29 (100)	0 (0)	0 (0)	0 (0)
Analysis and interpretation of results				
22. Time horizon of costs and benefits is stated.	26 (90)	3 (10)	0 (0)	0 (0)
23. The discount rate(s) is stated.	25 (86)	4 (14)	0 (0)	0 (0)
24. The choice of discount rate(s) is justified.	8 (28)	18 (62)	0 (0)	3 (10)
25. An explanation is given if costs and benefits are not discounted.	4 (14)	4 (14)	0 (0)	21 (72)
26. Details of statistical tests and confidence intervals are given for stochastic data.	19 (65)	10 (34)	0 (0)	0 (0)
27. The approach to sensitivity analysis is given.	28 (97)	1 (3)	0 (0)	0 (0)
28. The choice of variables for sensitivity analysis is justified.	26 (90)	3 (10)	0 (0)	0 (0)
29. The ranges over which the variables are varied are justified.	20 (69)	9 (31)	0 (0)	0 (0)
30. Relevant alternatives are compared.	22 (76)	0 (0)	7 (24)	0 (0)
31. Incremental analysis is reported.	18 (62)	4 (14)	7 (24)	0 (0)
32. Major outcomes are presented in a disaggregated as well as aggregated form.	21 (72)	8 (28)	0 (0)	
33. The answer to the study question is given.	21 (72)	3 (10)	5 (17)	
34. Conclusions follow from the data reported.	20 (69)	0 (0)	9 (31)	
35. Conclusions are accompanied by the appropriate caveats.	28 (97)	1 (3)	0 (0)	

and secondly, sufficient presentation of the ICER in disaggregated form to allow for correct interpretation. The ICER is a measure of cost-effectiveness when comparing two or more alternatives. It is calculated by dividing the difference in costs by the difference in effects. The new intervention is either better and more expensive, worse and less expensive, worse and more expensive, or better and cheaper. The two latter outcomes always predominate, and the first two outcomes have to be deemed cost-effective against a willingness-to-pay or willingness-to-accept threshold, respectively. Hence, a clearly stated ICER and transparent disaggregated presentation of results is vital to draw such logical conclusions, which was missing in eight studies [36]. Thirdly, the use of proxy measures raises some concerns. An example is the study by Hassan et al. which uses detection rates as outcome measure in the CEA [44]. However, such approach raises some challenges since not all adenomas will progress to cancer. This may well overestimate the effect of the technology, and in turn underestimate the ICER, which results in overutilization of scarce health care resources. All the abovementioned can suggest a more general lack of understanding of the important tenet of conducting high-quality HEEs, which is also evident in studies evaluating pharmaceuticals [50].

One limitation of the checklists used in this study is that they do not provide information on how the intervention impacts the outcome(s). AI interventions can be complex and often comprise several interactions between technology, healthcare professionals, and patients. It is the authors' opinion that there is a lack of understanding of how AI technologies impact the outcome(s). Lack of detailed analysis of costs is another limitation of the checklists. In general, there is a lack of understanding which cost should be included in a CEA and which cost are cost-drivers. A separate framework for cost and resource use is warranted.

None of the studies reported information on the learning curve, which is concerning since AI technologies can modify the effect as they mature and more data becomes available, altering cost-effectiveness. However, the authors acknowledges that this can be a very complex task to handle in practice [7, 15]. An AI-specific checklist, similar to the extensions of CONSORT-AI and SPIRIT-AI, could provide a more transparent quality assessment. Such checklist should also address which resources are relevant to include in CEA's. This is critical since AI carries various external costs, such as incremental innovation costs, costs related to re-training the model, monitoring costs etc.

We also recognise that AI is a generic term and that a one-size-fits-all approach will be difficult to apply given the distinctive differences between AI sub-types, such as purpose, technology, and effect. However, it is unavoidable that AI will impact future healthcare systems, given its ability to rethink healthcare by transforming large amounts of data to support diagnostics or clinical decision-making. Therefore, it is important to map possible methodological quality challenges in HEEs to enhance awareness of these in future research. Given these limitations, the list presented in this study provides a minimal level of guidance and is not sufficient for a full HEE for AI interventions. Thus, a more thorough checklist for HEE of an AI intervention is warranted. Regardless, only a few of the included studies adhere to this minimal level of guidance and can certainly be assessed as low quality HEEs.

The current study may be subject to bias as AI covers a broad terminology and research area, and our search strategy may not cover the entire field. However, we expect the number of false negatives is limited, and that our sample is representative for the current literature.

Conclusions

Given the large amount of application of AI interventions in healthcare there is a worryingly low number of full HEEs relative to the number of AI publications. Furthermore, mapping the evidence for the methodological quality of HEEs on AI shows a need to improve the quality specially concerning reporting and transparency of the ICER, and use of proxy outcome measures.

Abbreviations

ACL	Alongside clinical trial
AF	Atrial fibrillation
AI	Artificial intelligence
AI-ECG	Artificial intelligence electrocardiogram
ANN	Artificial Neural Networks
ARIAS	automated retinal image analysis system
CBA	Cost-benefit analysis
CEA	Cost-effectiveness analysis
CMAI-SF	Cohen–mansfield agitation inventory-short form
CPAP	Continuous positive airway pressure
CT	Cannot tell
CUA	Cost-utility analysis
CXR	Chest x-ray
DALY	Disability-adjusted life year
DAM	Decision analytic modelling
DR	Diabetic retinopathy
ECP	Eyecare professional
HEE	Health economic evaluations
ICE plane	Incremental cost-effectiveness plane
ICER	Incremental cost-effectiveness ratio
MeSH	Medical subject headings
NHS	National health service
OUDTEST	Opioid use disorder test
PHQ-9	Patient health questionnaire
PRISMA	Preferred reporting items for systematic reviews and meta-analyses-guidelines
PSA	Probabilistic sensitivity analysis
QALY	Quality-adjusted life year
ROP	Premature retinopathy
TB	Tuberculosis
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TTP	True positive proportion

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s44247-024-00088-7.

Supplementary Material 1.

Supplementary Material 2.

Supplementary Material 3.

Acknowledgements

We would like to thank Jette F. Jepsen from the medical library at Aalborg University Hospital for her valuable assistance with literature search.

Authors' contributions

NK have conceived the study idea. NK, AWHK, and JVB have made substantial contributions to the design of the work, analysis, and interpretation of results. All authors have reviewed and approved the final submitted version.

Authors' information

Not applicable.

Funding

NK has received internal unrestricted funding by Aalborg University, Aalborg University Hospital, and external unrestricted funding from Helsefonden (Grant 20-B-0008), and The Danish Ministry of Health. The funding sources will not have any influence on the research, authorship, and/or publication of results. The co-author(s) received no financial support for authorship, and/or publication of this article. We thank all the funding bodies for their vital financial support.

Availability of data and materials

The data collected and generated from the included studies can be sent on request to the corresponding author.

Declarations

Ethics approval and consent to participate

This article is based on previous studies and does not involve any new studies of human or animal subjects.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 20 February 2023 Accepted: 8 April 2024
Published online: 01 July 2024

References

1. Voets MM, Veltman J, Slump CH, Siesling S, Koffijberg H. Systematic Review of Health Economic Evaluations Focused on Artificial Intelligence in Healthcare: The Tortoise and the Cheetah. Value in Health. 2022;25:340–9.
2. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Watkinson P, et al. Consensus statement Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Mudathir Ibrahim. 12:28.
3. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. The BMJ. 2020;370:537–48.
4. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. Methods for the Economic Evaluation of Health Care Programmes. 4th ed. Oxford; 2015.
5. Chen Y. Health technology assessment and economic evaluation: Is it applicable for the traditional medicine? Integr Med Res. 2022;11:S516–7.
6. Drummond M, Griffin A, Tarricone R. Economic Evaluation for Devices and Drugs-Same or Different? Value in Health. 2009;12:402–4.
7. Drummond M, Tarricone R, Torbica A. Economic Evaluation of Medical Devices. In: Oxford Research Encyclopedia of Economics and Finance. Oxford University Press; 2018.
8. Davenport T, Kalakota R. Digital Technology The potential for artificial intelligence in healthcare. Future Healthc J. 2019;6(2):94–8.
9. Yousef Shaheen M. Article title: Applications of Artificial Intelligence (AI) in healthcare: A review Applications of Artificial Intelligence (AI) in healthcare: A review. 2021. https://doi.org/10.14293/S2199-1006.1.SOR-.PPVRY8K.v1.
10. Dhiman P, Ma J, Andaur Navarro CL, Speich B, Bullock G, Damen JAA, et al. Methodological conduct of prognostic prediction models developed using machine learning in oncology: a systematic review. BMC Med Res Methodol. 2022;22:101.
11. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? BMJ (Online). 2009;338:1317–20.
12. Faes L, Sim DA, van Smeden M, Held U, Bossuyt PM, Bachmann LM. Artificial intelligence and statistics: just the old wine in New Wineskins? Front Digit Health. 2022;4:833912.
13. Wolff J, Pauling J, Keck A, Baumbach J. The economic impact of artificial intelligence in health care: Systematic review. J Med Internet Res. 2020;22:16866.
14. Drummond MF. Education & Debate Guidelines for authors and peer reviewers of economic submissions to the BMJ. 1996;313(7052):275–83.
15. Tarricone R, Callea G, Ogorevc M, Prevolnik Rupel V. Improving the Methods for the Economic Evaluation of Medical Devices. Health Economics (United Kingdom). 2017;26:70–92.
16. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, The PRISMA, et al. statement: An updated guideline for reporting systematic reviews. The BMJ. 2020;2021:372.
17. van Mastrigt GAPG, Hiligsmann M, Arts JJC, Broos PH, Kleijnen J, Evers SMAA, et al. How to prepare a systematic review of economic evaluations for informing evidence-based healthcare decisions: a five-step approach (part 1/3). Expert Review of Pharmacoeconomics and Outcomes Research. 2016;16:689–704.
18. rayyan. rayyan.ai. https://www.rayyan.ai/. Accessed 6 Mar 2023.
19. Watts RD, Li IW. Use of Checklists in Reviews of Health Economic Evaluations, 2010 to 2018. Value in Health. 2019;22:377–82.
20. Walker DG, Wilson RF, Ritu Sharma M, John Bridges B, Niessen L, Bass EB, et al. Best Practices for Conducting Economic Evaluations in Health Care: A Systematic Review of Quality Assessment Tools. Rockville (MD): Agency for Healthcare Research and Quality (US); 2012.
21. Frederix GWJ. Check Your Checklist: The Danger of Over- and Underestimating the Quality of Economic Evaluations. Pharmacoecon Open. 2019;3:433–5.
22. Brandão M, Pondé N, Piccart-Gebhart M. Mammaprint^{TM}: a comprehensive review. Future Oncol. 2019;15:207–24.
23. Gomez Rossi J, Rojas-Perilla N, Krois J, Schwendicke F. Cost-effectiveness of Artificial Intelligence as a Decision-Support System Applied to the Detection and Grading of Melanoma, Dental Caries, and Diabetic Retinopathy. JAMA Netw Open. 2022. https://doi.org/10.1001/jamanetworkopen.2022.0269.
24. Delgadillo J, Ali S, Fleck K, Agnew C, Southgate A, Parkhouse L, et al. Stratified Care vs Stepped Care for Depression: A Cluster Randomized Clinical Trial. JAMA Psychiatry. 2022;79:101–8.
25. Fuller SD, Hu J, Liu JC, Gibson E, Gregory M, Kuo J, et al. Five-Year Cost-Effectiveness Modeling of Primary Care-Based, Nonmydriatic Automated Retinal Image Analysis Screening Among Low-Income Patients With Diabetes. J Diabetes Sci Technol. 2022;16:415–27.
26. Hill NR, Sandler B, Mokgokong R, Lister S, Ward T, Boyce R, et al. Cost-effectiveness of targeted screening for the identification of patients with atrial fibrillation: evaluation of a machine learning risk prediction algorithm. J Med Econ. 2020;23:386–93.
27. Huang XM, Yang BF, Zheng WL, Liu Q, Xiao F, Ouyang PW, et al. Cost-effectiveness of artificial intelligence screening for diabetic retinopathy in rural China. BMC Health Serv Res. 2022;22:260.
28. Mervin MC, Moyle W, Jones C, Murfield J, Draper B, Beattie E, et al. The Cost-Effectiveness of Using PARO, a Therapeutic Robotic Seal, to Reduce Agitation and Medication Use in Dementia: Findings from a Cluster-Randomized Controlled Trial. J Am Med Dir Assoc. 2018;19:619–622.e1.
29. Morrison SL, Dukhovny D, Chan RVP, Chiang MF, Campbell JP. Cost-effectiveness of Artificial Intelligence-Based Retinopathy of Prematurity Screening. JAMA Ophthalmol. 2022;140:401–9.
30. Schwendicke F, Mertens S, Cantu AG, Chaurasia A, Meyer-Lueckel H, Krois J. Cost-effectiveness of AI for caries detection: randomized trial. J Dent. 2022;119:104080.
31. 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.
32. 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.
33. Wolf RM, Channa R, Abramoff MD, Lehmann HP. Cost-effectiveness of Autonomous Point-of-Care Diabetic Retinopathy Screening for Pediatric Patients with Diabetes. JAMA Ophthalmol. 2020;138:1063–9.
34. Turino C, Benítez ID, Rafael-Palou X, Mayoral A, Lopera A, Pascual L, et al. Management and treatment of patients with obstructive sleep apnea using an intelligent monitoring system based on machine learning aiming to improve continuous positive airway pressure treatment compliance: randomized controlled trial. J Med Internet Res. 2021;23:24072.

35. Lin S, Ma Y, Xu Y, Lu L, He J, Zhu J, et al. Artificial intelligence in community-based diabetic retinopathy telemedicine screening in urban China: cost-effectiveness and cost-utility analyses with real-world data. JMIR Public Health Surveill. 2023;9:41624.
36. Chawla H, Uhr JH, Williams JS, Reinoso MA, Weiss JS. Economic Evaluation of Artificial Intelligence Systems Versus Manual Screening for Diabetic Retinopathy in the United States. Ophthalmic Surg Lasers Imaging Retina. 2023;54:272–80.
37. Barkun AN, von Renteln D, Sadri H. Cost-effectiveness of Artificial Intelligence-Aided Colonoscopy for Adenoma Detection in Colon Cancer Screening. J Can Assoc Gastroenterol. 2023;6:97–105.
38. Pickhardt PJ, Correale L, Hassan C. AI-based opportunistic CT screening of incidental cardiovascular disease, osteoporosis, and sarcopenia: cost-effectiveness analysis. Abdominal Radiology. 2023;48:1181–98.
39. Shen M, Zou Z, Bao H, Fairley CK, Canfell K, Ong JJ, et al. Cost-effectiveness of artificial intelligence-assisted liquid-based cytology testing for cervical cancer screening in China. Lancet Reg Health West Pac. 2023;34:100726.
40. Yonazu S, Ozawa T, Nakanishi T, Ochiai K, Shibata J, Osawa H, et al. Cost-effectiveness analysis of the artificial intelligence diagnosis support system for early gastric cancers. DEN Open. 2024;4:289.
41. Skarping I, Nilsson K, Dihge L, Fridhammar A, Ohlsson M, Huss L, et al. The implementation of a noninvasive lymph node staging (NILS) preoperative prediction model is cost effective in primary breast cancer. Breast Cancer Res Treat. 2022;194:577–86.
42. Mital S, Nguyen HV. Cost-effectiveness of using artificial intelligence versus polygenic risk score to guide breast cancer screening. BMC Cancer. 2022;22:501.
43. Srisubat A, Kittrongsiri K, Sangroongruangsri S, Khemvaranan C, Shreibati JB, Ching J, et al. Cost-Utility Analysis of Deep Learning and Trained Human Graders for Diabetic Retinopathy Screening in a Nationwide Program. Ophthalmol Ther. 2023;12:1339–57.
44. Hassan C, Povero M, Pradelli L, Spadaccini M, Repici A. Cost-utility analysis of real-time artificial intelligent-assisted colonoscopy in Italy. Endosc Int Open. 2023. https://doi.org/10.1055/a-2136-3428.
45. Hill NR, Groves L, Dickerson C, Boyce R, Lawton S, Hurst M, et al. Identification of undiagnosed atrial fibrillation using a machine learning risk prediction algorithm and diagnostic testing (PULsE-AI) in primary care: cost-effectiveness of a screening strategy evaluated in a randomized controlled trial in England. J Med Econ. 2022;25:974–83.
46. Ziegelmayer S, Graf M, Makowski M, Gawlitza J, Gassert F. Cost-effectiveness of artificial intelligence support in computed tomography-based lung cancer screening. Cancers (Basel). 2022;14:1729.
47. Nsengiyumva NP, Hussain H, Oxlade O, Majidulla A, Nazish A, Khan AJ, et al. Triage of persons with tuberculosis symptoms using artificial intelligence-based chest radiograph interpretation: a cost-effectiveness analysis. Open Forum Infect Dis. 2021;8:567.
48. Padula WV, Pronovost PJ, Makic MBF, Wald HL, Moran D, Mishra MK, et al. Value of hospital resources for effective pressure injury prevention: A cost-effectiveness analysis. BMJ Qual Saf. 2019;28:132–41.
49. Mallow PJ, Belk KW. Cost-utility analysis of single nucleotide polymorphism panel-based machine learning algorithm to predict risk of opioid use disorder. J Comp Eff Res. 2021;10:1349–61.
50. Erfani P, Bhangdia K, Stauber C, Mugunga JC, Pace LE, Fadelu T. Economic Evaluations of Breast Cancer Care in Low- and Middle-Income Countries: A Scoping Review. Oncologist. 2021;26:e1406–17.

Publisher's Note