Health and Technology (2023) 13:557–570
https://doi.org/10.1007/s12553-023-00763-1

REVIEW PAPER

Systematic review of the performance evaluation of clinicians with or without the aid of machine learning clinical decision support system

Mikko Nuutinen[1,2] · Riikka-Leena Leskelä[1]

Received: 13 February 2023 / Accepted: 1 June 2023 / Published online: 13 June 2023

Abstract

Background For the adoption of machine learning clinical decision support systems (ML-CDSS) it is critical to understand the performance aid of the ML-CDSS. However, it is not trivial, how the performance aid should be evaluated. To design reliable performance evaluation study, both the knowledge from the practical framework of experimental study design and the understanding of domain specific design factors are required.

Objective The aim of this review study was to form a practical framework and identify key design factors for experimental design in evaluating the performance of clinicians with or without the aid of ML-CDSS.

Methods The study was based on published ML-CDSS performance evaluation studies. We systematically searched articles published between January 2016 and December 2022. From the articles we collected a set of design factors. Only the articles comparing the performance of clinicians with or without the aid of ML-CDSS using experimental study methods were considered.

Results The identified key design factors for the practical framework of ML-CDSS experimental study design were performance measures, user interface, ground truth data and the selection of samples and participants. In addition, we identified the importance of randomization, crossover design and training and practice rounds. Previous studies had shortcomings in the rationale and documentation of choices regarding the number of participants and the duration of the experiment.

Conclusion The design factors of ML-CDSS experimental study are interdependent and all factors must be considered in individual choices.

Keywords Clinical decision support systems · Experimental design · Machine learning · Literature review

1 Introduction

A clinical Decision Support System (CDSS) is a software device that supports clinicians in decision making. For example, a CDSS can indicate areas on an X-ray image from where a fracture can be found [1, 2], guide an endoscopist to execute examination comprehensively [3] or warn clinicians of hypoxaemia/hypotension risk [4, 5].

Recent CDSSs utilize advanced machine learning (ML) techniques. However, traditional accuracy measurements of ML algorithms are not sufficient to show that the ML-CDSS is effective also in a real clinical environment. The human decision-making process is complex and biased. It cannot be assumed that clinicians will always closely follow the recommendations of ML models [6, 7]. For that reason, it is especially important to measure the performance of ML-CDSS software being developed and to validate the functionality well in advance by using suitable experimental methods before large-scale and expensive implementation.

In this study, we reviewed recent studies in which the performance of ML-CDSSs were measured using experimental study methods. The objective was to review how experimental studies measuring the performance of clinicians with or without the aid of ML-CDSS have been designed and conducted. Highlighting what aspects have been considered and what choices made in previous studies can provide guidance for the design of future experiments. Also, shortcomings in existing studies are identified, and places for improvement can be shown. For the review, we group the design factors and form a framework of ML-CDSS experimental

✉ Mikko Nuutinen
mikko.nuutinen@nhg.fi

[1] Nordic Healthcare Group, Helsinki, Finland

[2] Haartman Institute, University of Helsinki, Helsinki, Finland

study. In particular, we identify and explore the important individual ML-CDSS domain specific factors that require special attention.

The contribution of this study is that the studies selected for this review compare the performance of clinicians with and without the aid of a ML-CDSS. As far as we know, the studies selected for the previous review studies have compared mainly the performance of clinicians and ML models alone [8–11]. Furthermore, in this study we focus on the practical implementation of a ML-CDSS performance evaluation study. For example, recent artificial intelligence extensions [12, 13] for the clinical trial protocols and reporting guidelines focus on defining the items that should be reported, such as algorithm version and input/output data, not the questions about the practical implementation of experiments.

This review study is divided into two parts. Section 2 presents the methods used to search the published ML-CDSS performance evaluation studies and summarizes the selected studies (Table 1). Section 3 forms a framework of ML-CDSS performance evaluation study and groups the factors (Fig. 1), and discusses the important individual design factors.

Table 1 Publication year, disease, machine learning technology and expertise of the participants from the studies selected for the review. Diseases: number of cancer ML-CDSS studies: 14 (48.28%); number of fracture ML-CDSS studies: 4 (13.79%); number of other studies: 9 (31.03%); ML technologies: number of CNN techniques: 22 (75.86%); number of RL techniques: 1 (3.45%); number of other techniques: 7 (24.14%); Expertise: number of Anaesthesiologist: 2 (6.90%); number of pathologists: 2 (6.90%); number of endoscopist 4 (13.79%); number of radiologists: 16 (55.17%); number of others: 7 (24.14%). ML = Machine Learning; CNN = Convolution Neural Network; RL = Reinforcement Learning; GMM = Gaussian Mixture Model

Author	Year of publication	Disease	ML technology	Expertise of participants
Dhombres et al. [22]	2019	Pregnancy location	Knowledge based ontology	Obstetrics, gynecology
Lundberg et al. [4]	2018	Hypoxaemia	Gradient boosting	Anaesthesiologist
Steiner et al. [23]	2018	Breast cancer	CNN	Pathologist
Lindsay et al. [1]	2018	Wrist fracture	CNN	Emergency medicine clinician
Kaini et al. [24]	2020	Liver cancer	CNN	Pathologist
Wu et al. [3]	2019	Gastric cancer	CNN, RL	Endoscopist
Wang et al. [25]	2019	Colorectal cancer	CNN	Endoscopist
Bien et al. [26]	2018	Knee injury	CNN + Logistic regression	Radiologist, Orthopedic surgeon
Wijnberge et al. [5]	2020	Hypotension	Logistic regression	Anaesthesiologist
Su et al. [27]	2019	Colorectal cancer	CNN	Endoscopist
Zhou et al. [2]	2020	Rib fracture	CNN	Radiologist
Tajmir et al. [28]	2019	Bone age assessment	CNN	Radiologist
Sim et al. [29]	2019	Lung cancer	CNN, commercial tool	Radiologist
Lee et al. [30]	2020	Thyroid cancer	CNN	Radiologist
Kozuka et al. [31]	2020	Lung cancer	CNN, commercial tool	Radiologist
Jang et al. [32]	2020	Lung cancer	CNN, commercial tool	Radiologist
Cha et al [33]	2019	Muscle-invasive bladder cancer	CNN	Radiologist
Cai et al. [34]	2019	Esophageal cancer	CNN	Endoscopist
Sato et al. [35]	2021	Hip fractures	CNN	Clinician
Yu et al. [36]	2019	Breast cancer	GMM, Random forest	Radiologist
Choi et al. [37]	2021	Thoracic disease	CNN	Radiologist
Choi et al. [38]	2022	Skull fracture	CNN	Radiologist
Shang et al. [39]	2022	SARS-COV-2	CNN	Radiologist
Roller et al. [40]	2022	Graft failure	Gradient boosting	Physicians (internal medicine or nephrology)
Wang et al. [41]	2022	Pancreatic cancer	CNN	Radiologist
Yacoub et al. [42]	2022	Cardiac, pulmonary and musculoskeletal diseases	CNN, commercial tool	Radiologist
Wei et al. [43]	2022	Breast cancer	Commercial tool	Radiologist
Wataya et al. [44]	2022	Lung cancer	CNN	Radiologist
Toda et al. [45]	2022	Lung cancer	CNN, commercial tool	Radiologist, pulmonologist

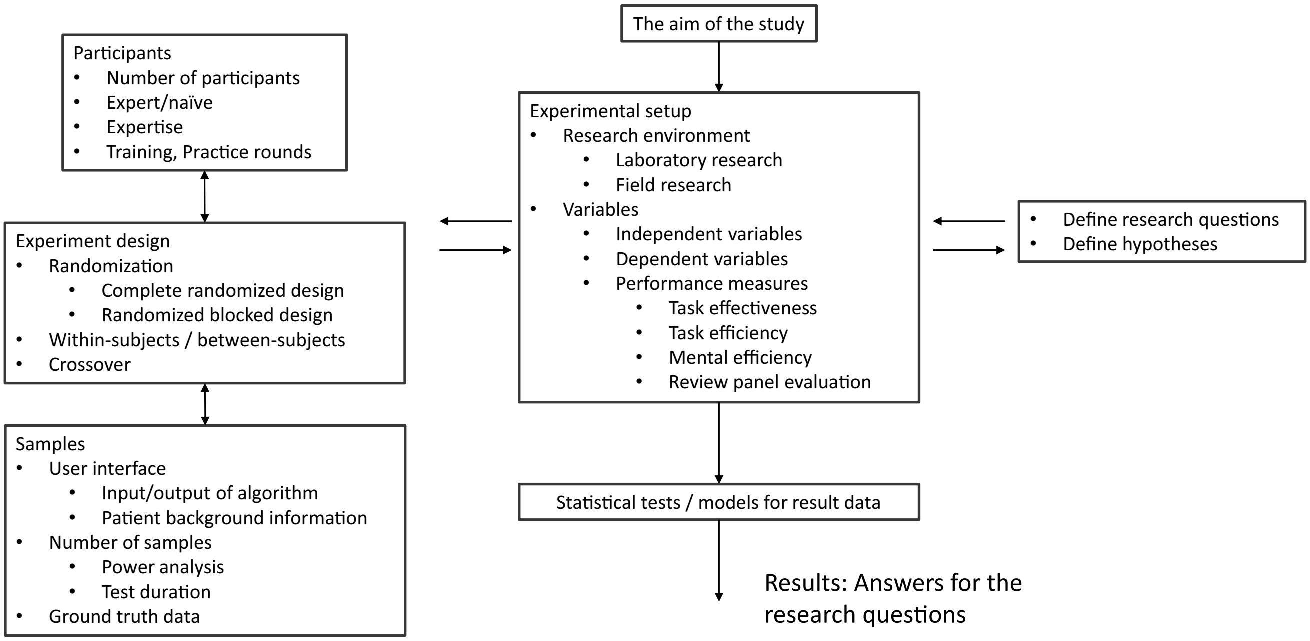

Fig. 1 The framework of ML-CDSS performance evaluation study. The factors of the framework are grouped into five groups: experimental setup, participants, experiment design, samples and statistical tests and models for result data

2 Methods

2.1 Search strategies

Our literature search was conducted in PubMed using the combination of search terms (see Appendix). The search was limited to articles published between January 2016 and December 2022.

We included articles if study compared the performance between clinicians with the aid and without the aid of a CDSS, CDSS was based on ML techniques (ML-CDSS), and study was experimental and systematically designed. Articles were excluded if the comparison was between the performance of ML algorithm alone and the performance of a clinician without the CDSS (e.g. [14–19]), or study was observational (e.g. [20, 21]).

On the basis of the above inclusion and exclusion criteria, one author (MN) screened article titles and abstracts and identified eligible articles. The full texts of eligible articles were retrieved. Indistinct samples of this process were resolved by discussion with other authors.

2.2 Results

Following the search process, 1276 citations were retrieved from the database and 1152 articles were excluded based on their titles and abstracts, resulting in 124 articles to be reviewed in detail. In addition, 94 articles were further excluded based on their full text. Finally, 29 studies were included for review.

The data we obtained from each study were year of publication, disease, ML technology, research type (laboratory/field), independent and dependent variables, performance measures, participants (number of subjects, expertise, training), experiment design (randomization, crossover), samples (user interface, number of samples), test duration, ground truth data and performance values.

Table 1 presents top-level figures, such as year of publication, disease, ML technology and expertise of participants. All the selected studies were published between 2018-2022. In many studies, the ML-CDSS was developed to help diagnose cancer (48.28% of the studies). ML technology was based in almost all studies on convolutional neural networks (75.86% of studies). The participants of the studies were most often radiologists (55.17% of the studies).

3 Framework for clinical DSS performance evaluation study

Our analysis follows the framework depicted in Fig. 1. The framework, derived from research literature [46, 47], groups the factors of ML-CDSS performance evaluations study in five groups: experimental setup (Section 3.1), participants (Section 3.2), experiment design (Section 3.4), samples (Section 3.3) and statistical tests and models for result data. Next, Sections 3.1–3.4 discusses individual factors of each group.

3.1 Experimental setup

The objective of study defines frames for experimental setup. That is, what is measured (variables that are modified) and where experiment is conducted (research environment). When research environment and independent and dependent variables are fixed, research questions and hypothesis can be refined.

3.1.1 Research environment

Research environment can be classified as a field study or a controlled laboratory study or something between them. A field study refers to a natural clinical environment where clinicians use the ML-CDSS. With field studies, not all variables can be controlled for and their effects should be taken into account in the study design. A controlled laboratory study is carried out in an isolated space and time. In laboratory studies, important research variables are modified while other variables are constant.

In the studies in this review, the distinction between a laboratory and a field study was made based on the data type. If a study used retrospective data and participants used ML-CDSS for simulating patient examination, it was considered laboratory research. If a study collected prospective data and ML-CDSS was used as a part of normal patient examination, it was considered field research. Five studies in this review (17.24% of the studies) were classified as field studies (see Table S1 in Supplementary material). In studies [3, 5, 25, 27, 42], clinicians examined consecutive patients randomly with the aid of ML-CDSS or without the aid. The other 25 studies were laboratory studies. These studies used retrospective data that was presented to participants who executed tasks with the aid of ML-CDSS or without the aid. The tasks of these studies simulated patient examinations.

3.1.2 Independent and dependent variables

An experimental performance study measures the change of the dependent variable when the level of independent variable is changed. In the studies in this review, the primary independent variable was ML-CDSS aid, which has the two levels: with and without the aid. Dependent variable of the studies was, for example, the number of detected findings, diagnoses estimated by participants or reaction/review time. The primary hypothesis was that the aid of ML-CDSS increases the performance in executing tasks. That is, with the aid of the ML-CDSS, clinicians estimate diagnoses more accurately or detect symptoms faster than without the aid of the ML-CDSS.

Some studies in this review (51.72% of the studies) had more than one independent variable (see Table S1 in Supplementary material). For example, study [1] grouped the participants in the groups of medical doctors and physician assistant. Study [24] grouped the participants in the groups of GI (gastrointestinal) subspecialty, non-GI subspecialty, trainee and pathologist not-otherwise classified. Study [29] grouped the participants in the groups of resident or chest radiologists. Study [30] groups the participants in the groups of trainee or staff radiologist. Study [24] measured the effect of tumor grade for the clinicians' performance. Study [33] measured the effect of the difficulty of chemotherapy response evaluation for the clinicians' performance. These additional independent variables answered to the research question if clinician's professional level or the difficulty of the task had an effect on the clinician's performance to execute tasks with or without the aid.

3.1.3 Performance measures

If participants perform tasks more effectively or efficiently at one level of the independent variable than at the other, the value of the dependent variable changes. The change is quantized by performance measures. Study [48] grouped the performance measures in three categories: task effectiveness, task efficiency and mental efficiency. In addition to this, we defined a fourth category: review panel evaluation. Review panel evaluation measures are values that are scored after the study by an external expert panel.

Table 2 groups the dependent variables and performance measures of the studies in this review into four performance measure categories. 24 studies (82.76% of the studies) in this review, used task effectiveness measures (see Table S1 in Supplementary material). Task effectiveness measures indicate the ratio of correctly classified or evaluated samples to total number of samples. These measures identify whether the ML-CDSS assisted participants to perform tasks more effectively. For example, the accuracy of diagnoses (e.g., early pregnancy, cancer diagnose, knee injury, wrist or rib fractures, bone age) was an often-used performance measure [22, 24, 26, 28–30]. Also, the performance measures of sensitivity and specificity [1, 2, 23, 26, 31, 41, 43, 44] and AUC value (area under curve) [4, 32, 33, 37, 38, 40, 43–45] were often used.

11 studies (37.93% of the studies) in this review, used task efficiency measures (see Table S1 in Supplemenary material). Task efficiency measures indicate if ML-CDSS aided clinician to execute tasks more efficiently. These measures are related to examination time or the number of repetitions, detections or incidents. For example, studies [2, 3, 22, 23, 25, 27, 32, 41, 42, 44] measured time to execute tasks (e.g., scan duration, review time, reaction time, reading time). Studies [3, 5, 22, 25] measured the numbers of detections (e.g., number of polyps, number of unobserved sites), repetitions (e.g., number of scan images) and incidents (e.g., number of treatments or hypotensive events).

Table 2 Performance measures and dependent variables from the studies of this review for evaluating clinicians with or without the aid of ML-CDSS were categorized into the groups of task effectiveness, task efficiency, mental efficiency and review panel evaluation. AUC = Area Under Curve, RMSE = Root Mean Square Error

	Performance measure	Dependent variable	Authors
Task effectiveness	Accuracy, sensitivity, specificity	Pregnancy location and diagnosis	[22]
		Cancer diagnosis/detection	[23, 24, 29–32, 34, 36, 41, 43, 44]
		Knee injury diagnosis	[26]
		Fracture diagnosis	[1, 2, 35, 38]
		Bone age (one year range)	[28]
		SARS-CoV-2 diagnosis	[39]
		Abnormal finding	[45]
	AUC	A relative risk of hypoxaemia	[4]
		Confidence rate if malignant lung lesion is detected	[32, 45]
		Likelihood of complete response of chemotherapy	[33]
		Localizing thoracic abnormalities	[37]
		Likelihood of fracture	[38]
		Likelihood of graft failure	[40]
		Likelihood of presence of the 15 characteristics (pulmonary nodules nodules)	[44]
		Dichotomized pattern of benign or malignant breast masses	[43]
	RMSE	Bone age	[28]
Task efficiency	Time (avg)	Reading, review, withdrawal, scan, inspection, diagnosis time	[1–3, 5, 22, 23, 25, 27, 32, 41, 42, 44]
		Time-weighted average of hypotension, total time with hypotension, percentage of time spent with hypotension during surgery	[5]
	Number of repetitions	Number of scans	[22]
	Number of detections	Number of detected polyps/adenomas	[25, 27]
		Number of unobserved sites in patient	[3]
		Number of chest CT recommendations	[32]
	Number of incidents	Number of hypotensive events per patient	[5]
		Number of treatments per patient	[5]
Mental efficiency	Obviousness score	Obviousness rate of breast cancer lymph node	[23]
	Confidence score	Level of confidence in identifying cervical LNM	[30]
		Confidence rate for detected malignant lung lesion	[32]
Review panel evaluation	Trust	Trust score (results of simulated ultrasound imaging)	[22]
	Quality	Quality of image set (results of simulated ultrasound imaging)	[22]
	Completeness	Completeness of photo documentation (real time esophagogastroduodenoscopy)	[3]

Detection rate is a typical value derived from the numbers of detections. For example, study [25] calculated the detection rates of polyps/adenomas per sample. That is, how many polyps/adenomas clinicians found from one colonoscopy patient with the aid of ML-CDSS or without the aid.

Mental efficiency measures indicate required mental resource or confidence to perform a task. Mental efficiency can be measured with a self-reported questionnaire, which collects numerical values or comments. Three studies in this review [23, 30, 32] measured mental efficiency values (see Table S1 in Supplementary material). Study [23] used the scale of 0-100 to rate obviousness, when the task was to decide the category of negative, isolated tumor cells, micrometastasis, or macrometastasis from digitized slides from lymph node sections. Study [30] used the scale of 1-5 when the task was to measure the level of confidence in identifying cervical lymph node metastasis. Study [32] used the scale of 1-100 when the task was to measure the level of confidence in detecting potential malignant lung lesions.

Review panel evaluation relates to an external expert panel which performs a post-hoc evaluation for the results/documentation of a task. Two studies in this review used review panel evaluation measures (see Table S1 in Supplementary material). For example, in study [22] an external expert panel evaluated the trustworthiness of documentation produced by the participants of an experiment. The task of the participants was to diagnose early pregnancy and pregnancy location from ultrasound imaging. In study [3], an external expert panel evaluated the completeness of photo documentation produced by the participants who conducted esophagogastroduodenoscopy examinations.

3.1.4 Statistical tests

Statistical tests are used to prove the statistical significance of the difference between the values of performance measure for the different levels of the independent variable. The selected performance measure defines the requirements for statistical tests. Different tests are used for continuous, category and count data. In many studies, t-test was used because the output of performance measures was continuous value [4, 22, 26, 27, 39–42]. Also, non-parametric Mann Whitneu U test was used in some studies [3, 44]. For the categorical outputs, for example, the exact McNemar or chi-squared tests were used [3, 5, 22, 26, 27, 41–43].

3.2 Participants

With traditional randomized clinical trials, the patients undergoing medical examination are the participants of the study. With ML-CDSS performance experiment studies, clinicians who execute experimental tasks are the participants of the study.

In general, experimental study participants are classified as naïve or experts. Naïve participants do not have deep understanding or experience in the domain whereas expert participants do. The participants in the all studies in this review were experts (Table 1). They were pathologists [23, 24], endoscopists [3, 25, 27], radiologists [26, 28–33, 36, 37, 41–44], orthopedic surgeons [26], internal medicine and nephrology physicians [40], emergency physicians [38] or anaesthesilogists [4, 5].

The important study design question is the number of participants. In the studies reviewed, the number of participants was mainly between 2-16 (see Table S2 in Supplementary material). Studies [1] and [35] were exceptions with 40 and 31 participants. According to recommendations, the number of naïve participants should be more than 15 [49] or 20 [50]. Statistically significant results are possible to achieve with lower numbers of expert participants than with naïve participants. According to the study [51], the number of expert participants should be 10-15. However, none of the studies in this review discussed how they chose the number of expert participants.

3.3 Samples

3.3.1 User interface

All data (patient information) of the experiment is presented via the user interface (UI) for participants and the values of dependent variables are entered using the input elements of the UI. We identified two patient information types presented in the UI: patient background information and decision support information. Patient background information means patient-related medical knowledge, such as medical history and results from physical examinations, laboratory or imaging findings. Decision support information means the output information of the ML algorithm incorporated in the ML-CDSS.

Table 3 presents patient background and decision support information that was presented for participants in the studies in this review. We found that 22 studies in this review presented only decision support information on UI, such as heat maps or bounding boxes on medical image [1, 2, 22–24, 26, 29–39, 42–45], but no patient background information on the UI. Only seven studies [3–5, 25, 27, 40, 41] presented patient background information on the UI.

We further divided decision support information into two types:

- Category support information: a predicted probability/category of diagnosis/state of patient
- Guidance support information: an instructional guidance for participants to execute patient examination comprehensively or better.

In 25 studies in this review (86.2% of the studies), decision support information belonged to the category group (Table 3). In many of these cases, decision support information was presented as heat maps on medical images. For example, in study[1] if the ML algorithm found a fracture, heat map was used for showing the location of the fracture and the confidence of the model's prediction. Study [23] visualized confidence that tissue contains tumor by using cyan and green rectangles on images. Study [26] highlighted regions on an image that were important for the model's knee injury classification decision.

In four studies [3, 22, 25, 27], decision support information belonged to the guidance group (Table 3). The aid tested in these studies was instructional guidance for participants to execute a task more efficiency or comprehensive. For example, in study [3] a virtual stomach model was presented to guide endoscopist to find blind spots. In studies [25, 27] bounding boxes were presented on the video image for showing the locations of polyps. Furthermore, audio prompts were played to help tune withdrawal speed or to alarm for potential polyps. In study [22] the ML-CDSS presented keyword suggestions for participants for selecting reference images from a database.

Table 3 The patient background information and decision support information presented on the ML-CDSS user interface. Patient background information means patient-related medical knowledge, such as medical history and results from physical examinations, laboratory or imaging findings. Decision support information means the output information of the machine learning algorithm incorporated in the ML-CDSS. Decision support information is grouped into two types: Category support information and Guidance support information. Category support information presents a predicted probability/category of diagnosis/state of patient. Guidance support information presents an instructional guidance for participants to execute patient examination comprehensively or better

Author	Task	Patient background information	Decision support information	Decision support: Category/ Guidance
Dhombres et al. [22]	Ultrasound examination for early pregnancy diagnostic	N/A	Guided keywords selection, presentation of reference images for the selected keywords and suggestions of additional views	Guidance
Lundberg et al. [4]	Hypoxaemia risk prediction	Patient record data	Risk of hypoxaemia in the next five minutes and explanations	Category
Steiner et al. [23]	Lymph node image classification	N/A	Regions on the lymph node image were highlighted based on the algorithm predictions.	Category
Lindsey et al. [1]	Wrist fracture detection	N/A	Regions on the radiograph image were highlighted based on the algorithm predictions	Category
Kiani et al. [24]	Subtype of primary liver cancer classification	N/A	Regions on the image crop of H &E stained digital WSI were highlighted based on the algorithm predictions	Category
Wu et al. [3]	EGD study and photodocumentation	All patient information was available (field study)	Virtual stomach model for monitoring blind spots (sites not detected yet by the endoscopist)	Guidance
Wang et al. [25]	Colonoscopy examination for polyp and adenoma detection	All patient information was available (field study)	The detected polyps are presented with a hollow blue tracing box with a sound alarm	Guidance
Bien et al. [26]	Knee injury classification	N/A	Regions on the MRI were highlighted based on the algorithm predictions.	Category
Wijnberge et al. [5]	Elective noncardiac surgical procedure with the early warning system of intraoperative hypotension	All patient information was available (field study)	If the predicted risk of hypotension was high, sound and flickering light were presented. Mixture of variable information provided information about the underlying cause of the predicted hypotension.	Category
Su et al. [27]	Colonoscopy examination for polyp and adenoma detection	All patient information was available (field study)	(1) a real-time timer; (2) an audio prompt to remind to slow down the withdrawal speed and reexamine certain colonic segments when unstable or blurry frames were detected; (3) an audio prompt to persuade to clean the mucosa or suction liquid pools if needed; (4) a bounding box showing the location of a polyp	Guidance
Zhou et al. [2]	Rib fracture detection and classification	N/A	CT image with rectangular boxes indicating predicted fractures	Category
Tajmir et al. [28]	Bone age assessment	N/A	Bone age prediction and attention map on the radiograph	Category

Table 3 (continued)

Author	Task	Patient background information	Decision support information	Decision support: Category/ Guidance
Sim et al. [29]	Malignant pulmonary nodule detection from chest radiographs	N/A	Dotted circles were presented on the radiograph based on the algorithm prediction	Category
Lee et al. [30]	Cervical lymph node metastasis diagnosis from CT images	N/A	Regions on the CT image were highlighted based on the algorithm predictions	Category
Kozuka et al. [31]	Pulmonary nodule detection from CT images	N/A	Marks, density, major axis, and the volume of detected nodules on the CT image	Category
Jang et al. [32]	Pulmonary nodule detection from radiograph	N/A	Color coded map indicating a probability that a radiograph contains a malignant nodule	Category
Cha et al [33]	Chemotherapy response detection	N/A	Pre- and post-treatment CT scans were presented side-by-side with likelihood score of response	Category
Cai et al. [34]	Esophageal squamous cell carcinoma (ESCC) detection	N/A	The lesions on the image were marked with a square based on the algorithm predictions	Category
Sato et al. [35]	Hip fracture detection	N/A	Regions on the radiograph image were highlighted based on the algorithm predictions	Category
Yu et al. [36]	Breast lesions diagnostic	N/A	Regions on the radiograph image were highlighted based on the algorithm predictions	Category
Choi et al. [37]	Thoracic abnormalities detection from chest radiographs	N/A	Regions on the radiograph image were highlighted based on the algorithm predictions	Category
Choi et al. [38]	Skull fracture detection	N/A	Regions on the radiograph image were highlighted based on the algorithm predictions (+ probablity prediction)	Category
Shang et al. [39]	SARS-CoV-2 virus diagnosis from lung ultrasonography	N/A	Regions on the LUS image were highlighted based on the algorithm predictions (+ classification result)	Category
Roller et al. [40]	Detection of graft failure	Complete patient history, including text notes, medical reports, laboratory tests	Risk scores, relevant features which influence the decision of the risk score	Category
Wang et al. [41]	Pancreatic lesion diagnostic	Radiological characteristics	Risk prediction	Category
Yacoub et al. [42]	Chest CT interpratation	N/A	Analysis of finding, including labeling, segmenting and measuring normal structures as well as detecting, labeling and measuring abnormalities	Category
Wei et al. [43]	Breast lesions diagnostic	N/A	Risk prediction	Category
Wataya et al. [44]	Pulmonary nodule diagnostic from radiograph	N/A	Presents characterization results from lung nodule and three candidate radiology reports	Category
Toda et al. [45]	Detection of pulmonary nodules, masses and consolidation	N/A	Detects pulmonary nodules/masses and consolidation and marks the areas of the lesions	Category

3.3.2 Number of samples

In the studies in this review the term sample refers to an entity being examined by the participant, i.e. a patient or patient data, such as imaging data, vital signs or other patient record data. Many factors affect the required number of samples, such as evaluation time of one sample, availability of participants and the sensitivity of dependent variables for the changes of the independent variable.

Power calculation is a traditional method for defining the minimum number of samples required for statistically significant results. Seven studies [3, 5, 25, 27, 37, 40, 42] of this review used and reported power calculations (see Table S3 in Supplementary material). The numbers of samples required according to the power calculation were between 30 and 651.

If the number of samples is high and test duration becomes too long, participants tire, and the quality of collected data suffers. Also, new participants can be difficult to recruit for long experiments. Only one study [23] of this review discussed or documented test duration: the test duration was 3 hours including training, instructions, and breaks (see Table S3 in Supplementary material).

Also, training and practice rounds lengthen the duration of experiment. Although, the training and practice rounds are important for reliable results, only five studies in this review reported that they conducted some training before the experiment (see Table S3 in Supplementary material). Study [23] presented five and study [24] four training samples for the participants. Study [22] presented a 2 minutes video and conducted a 10 minutes hands on session before the experiment. In study [44], the participants received training on the definition of characteristics of the platform before the experiment. In study [42], the participants received training on interpreting the platform and used the platform for at least 30 days.

3.3.3 Ground truth data

One important study design question is how ground truth (GT) data is produced. GT data must be a close estimate for the true values of samples. In this study we divided GT data production methods into four types: (1) Majority vote, (2) High expertise, (3) Many data sources, and (4) Numerical data (see Table S3 in Supplementary material).

Majority vote method assumes that the majority opinion of a review group is GT for a sample. In many settings, majority vote and high expertise methods are used together. That is, the group of highly experienced clinicians is used for voting. For example, in study [23] GT data was produced by the majority vote of three experts (US board certified pathologists, > 7 years of experience). In that study and also in the studies [1–3, 26, 29, 31, 34, 45, 52], GT method was both majority vote and high expertise.

If the number of experts used for generating GT data is smaller than three, the method is classified as high expertise only, not majority vote. For example, in study [22] the documents produced by participants were reviewed and scored by two senior experienced ultrasound operators.

The method of deriving GT data from many data sources combines information from different sources that are not available for participants. These data sources are, e.g., patient records and other longitudinal patient tracking data. The method is viable in particular for retrospective studies where longitudinal patient data after patient examination is available. For example, in study [4] the hypoaxemia states of the patients retrieved from patient records were considered GT data. Other examples of the GT method of many data sources are studies [24, 26, 28, 30, 33, 34, 36, 37, 39, 40, 43]. In study [26], expert group produced GT data by using all DICOM series, clinical history and follow-up exams of samples. In study [28], experts had access to machine learning attention maps, machine learning bone age scores, and the clinical reports to define the bone age from radiographs.

Whereas usually GT data represents the "true diagnosis", GT method of numerical data refers to settings where the evaluation is not about the correctness of the finding, but rather number of findings or the speed of decision making. For example, in studies [25, 27] the performance was calculated based on the number of detected and analyzed polyps/adenomas. Higher number of detected polyps/adenomas was evaluated to be a better result. In study [5] shorter hypotension time of the patient and a shorter reaction time of the participants were better results. In study [42] shorter interpretation time of the chest computer tomography image was better result.

Post-hoc GT data means that ground truth values are produced after the experiment. That is, no prior true values for samples exist. For example, in study [3], after experiment, two seniors (1–5 years of experience) and three experts (>5 years of experience) reviewed and scored endoscopy videos and documentation generated by the participants. It should be noted, that the concepts of post-hoc GT and review panel evaluation measure group (Table 2) are similar or same.

3.4 Experiment design

3.4.1 Randomization and within-subjects/between-subject design

Traditional experiment design can be, e.g., complete randomized design (CRD) or randomized block design (RCBD). It should be noted that CRD and RCBD concepts differ somewhat between traditional clinical trials and ML-CDSS experiment studies. In the first case, patients are the participants of the study who are randomized into groups (yes/no treatment) forming the independent variable. In the latter case, clinicians are the participants of the study, and the independent variable to be randomized is with/without the aid of ML-CDSS.

The CRD divides clinicians randomly into two groups. One group executes tests with the aid of a ML-CDSS and the other group without it. In RCBD, first the heterogenous participant group is divided into homogeneous same-size sub-groups (blocks). Next, examinations with or without the aid are randomized for different blocks. In that way, the same number of participants from different blocks execute the experiment with and without the aid of a ML-CDSS. This can eliminate variance sources from confounding factors, such as professional level of participants. That is, RCBD ensures that with/without aid condition has, for example, an equal proportion of different professional levels. As a result, differences between the conditions cannot be attributed to professional level.

Furthermore, the experiment design can be classified as within-subject or between-subject design. In the between-subject design, one clinician conducts the experiment always at the same level of the independent variable. In the within-subject design, all levels of independent variable are presented for all clinicians. That is, between-subject design divides clinicians into two groups: one group executes tests with the aid and other group without the aid. Within-subject design presents the with aid and without aid settings for all clinicians. That is, all clinicians evaluate the same sample twice (with and without aid) or part of samples with the aid and part of samples without the aid. In this review, 82.8% of the studies applied within-subject experiment design (see Table S4 in Supplementary material). For example, studies [1, 22, 29, 31, 45] presented all samples without the aid first and then with the aid for all participants. Study [23] presented samples randomly with or without the aid first for all participants and then the opposite for the second round. Studies [2, 5, 28, 30, 43] presented one sample first without the aid of ML-CDSS and then with the aid.

3.4.2 Crossover

Crossover is an important concept for evaluating the performance of the aid of ML-CDSS. Crossover design executes two or more experiment sessions for each participant. For example, in the first experiment session, participants evaluate half of the samples with the aid and half of samples without the aid. Then, after a washout period, in the second session, participants evaluate the same samples but now with aid, if originally evaluated without aid, and vice versa. It is assumed that the washout period decreases the memory footprint from the first session. It should be noted, that crossover design is limited for laboratory studies in which retrospective data is used.

Twelve studies [22–24, 26, 29, 31, 32, 34, 38, 44] in this review (41.4% of the studies) used a crossover design (see Table S4 in Supplementary material). The length of the washout period varied. In studes [2, 22] the washout period was two months. In study [29] the washout period was only 2-6 hours. In other studies, the washout period was from 7 to 28 days. When planning the length of the washout period, it is important to note that participants may recall samples for a long period after the experiment, particularly the difficult samples. Study [53] recommended that the washout period should be at least 2 weeks. On the other hand, with a long washout period, the participant's diagnostic criteria could have changed over time. For example, participants could have gained more experience or changed their attitude toward diagnostic criteria [54].

4 Discussion

The aim of this review study was to summarize experimental study design factors for measuring the performance of clinicians with or without the aid of a ML-CDSS. The key factors were performance measures, user interface, ground truth data, samples and participants.

Figure 2 shows dependencies between the selection of proper performance measure and dependent variables and ground truth. First, dependent variables (input requested from the participant) determine which performance measures can be calculated. With the task effectiveness measures (such as accuracy, sensitivity, specificity and AUC), input value is binary [0,1] or a probability/continuous value. If the AUC value is the preferred performance measure, then the input should be a probability/continuous value. Accuracy, sensitivity and specificity performance measures can be calculated from binary inputs. It's important to note, that probability/continuous values can be harder for participants to estimate than binary states. Second, the research environment (retrospective vs. prospective study or laboratory vs. field research) also affects selectable performance measures. With prospective research settings, no task effectiveness measures (such as AUC, accuracy, sensitivity, specificity) can be used, because no exact numerical GT values are available. Task effectiveness measures require exact GT values, such as true diagnosis or patient state. That is, for the prospective research settings, the performance measures of task efficiency, mental efficiency or review panel evaluations should be used.

Different performance measures provide different information about the ML-CDSS. Task effectiveness and task efficiency measures calculate simple numerical or ratio values (e.g. accuracy level or detection ratio). If more deep or subjective evaluations, such as benefits or pitfalls of the system, are required, the use of mental efficiency measures

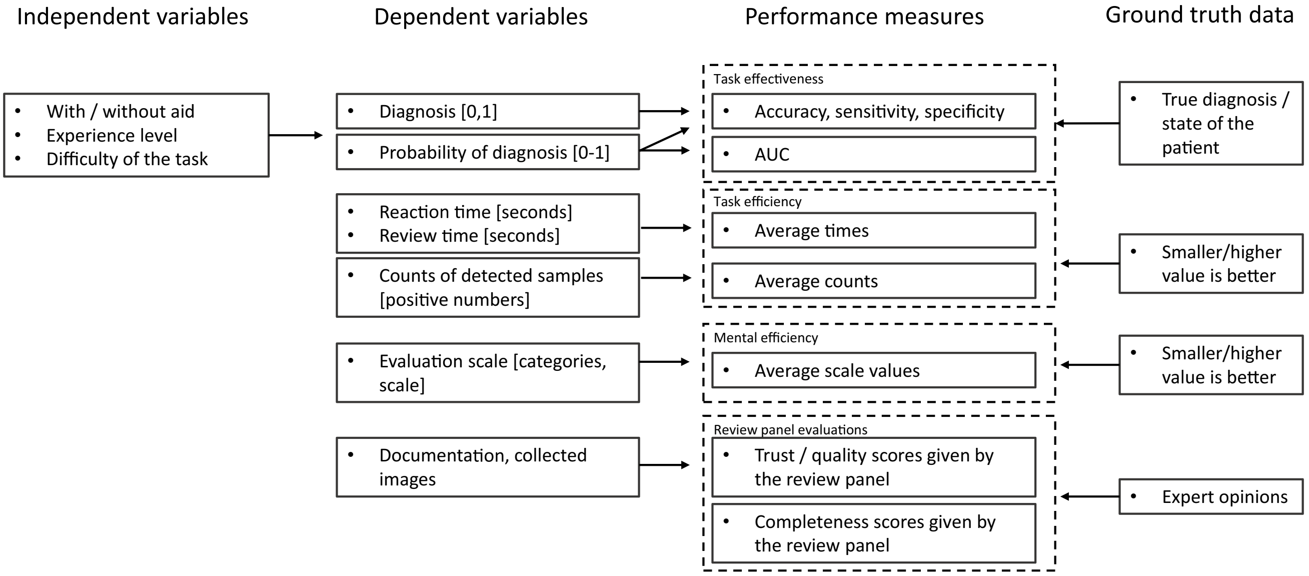
Independent variables
Dependent variables
Performance measures
Ground truth data
With / without aid
Experience level
Difficulty of the task
Diagnosis [0,1]
Probability of diagnosis [0-1]
Reaction time [seconds]
Review time [seconds]
Counts of detected samples [positive numbers]
Evaluation scale [categories, scale]
Documentation, collected images
Task effectiveness
Accuracy, sensitivity, specificity
AUC
Task efficiency
Average times
Average counts
Mental efficiency
Average scale values
Review panel evaluations
Trust / quality scores given by the review panel
Completeness scores given by the review panel
True diagnosis / state of the patient
Smaller/higher value is better
Smaller/higher value is better
Expert opinions

Fig. 2 Experimental performance study of clinical decision support system aims to prove that by changing the independent variable the dependent variables change. Performance measures are used to measure the significancy of the change. Performance measures are calculated by comparing the values of dependent variables and ground truth. In this study we grouped the performance measures into the groups of task effectiveness, task efficiency, mental efficiency and review panel evaluation. Different ground truth data types are used for different performance measures

should be considered. For example, mental efficiency measures can be used for evaluating the difficulty of the task or participant's confidence or what are situations when ML-CDSS interferes decision making. Answers can be given on a numerical scale or in open answers. For example, study [55] used a method of "thinking-aloud-test" in which the participants communicate their thoughts aloud while interacting with the system. For the situations when the outputs of participants are difficult to quantify, such as written reports or other documentation, the evaluation of review panel measures can be the proper choice. However, review panel evaluation measures are expensive to use, when the number of participants is high.

A key step of designing the elements of UI is to determine available patient information. In this review, we divided patient information into the groups of patient background and decision support. Patient background information is all relevant medical knowledge about the patients. Decision support information is based on the output of the incorporated ML algorithm. First step of designing the UI is to review information available at a normal patient examination. That is, patient background information that clinicians use for planning treatments and interventions. We found that many studies in this review presented only decision support information on UI, but no patient background information on the UI. It should be noted that the results of laboratory type studies are comparable to normal patient examinations only when all patient's background information relevant to the task is presented in the UI.

The design of the number of participants and samples and the design of duration of test are interdependent. The required number of samples depends on how large the difference is between the levels of dependent variable. If the difference is small, the number of samples should be higher for significant results. Power analysis can be used for approximating number of samples. The total number of samples can be increased either by increasing the number of participants or by increasing the number of samples evaluated by one participant. It should be remembered, though, that when the number of samples evaluated by one participant is increased, experiment duration lengthens. Too long an experiment causes fatigue, which lowers the quality of input values. It is also important to document all relevant information about the samples and participants. Only one study [23] in this review documented the duration of the experiment. One rule of thumb for determining the number of samples is to limit the duration of the test to 30 minutes. Then by calculating an average time to evaluate one sample, a maximum number of samples can be calculated. The number of participants should be high enough to produce significant differences between the samples.

In addition to the factors of experimental design, there are other important issues involved in the practical implementation of ML-CDSS experiments that should be recognized. For example, the accuracy of the ML method incorporated in ML-CDSS affects the performance of the clinicians. For example, in study [24], when the ML algorithm predicted correct diagnoses, ML-CDSS improved the accuracy of the clinicians. When the predictions were incorrect, ML-CDSS significantly decreased their accuracy.

The conclusions of the experiment should analyze whether the use of a ML-CDSS caused false negatives. That is, positive samples were missed because of the suggestions of the ML-CDSS. A reason can be that clinicians relied more on the ML-CDSS than their own conclusions. False negatives can be related to the low accuracy of ML algorithm, but may also be related to the information presented on UI. For example, if explanations for the factors affecting the predictions of the ML algorithm are presented and are interpretable in the UI, the clinician can interpret whether the prediction is reliable and whether it should be taken into account in the decision making. One important research question of ML-CDSS performance evaluations presented in recent studies is how providing explanations of ML model results and factors affecting it benefits decision making [48, 56, 57].

Finally, when drawing conclusions from the results of the experiment, it is important to keep in mind that experimental environment often does not correspond to a real clinical environment. For example, in a laboratory study of ML-CDSS, there are no unrelated distractions, nor are there other examinations requiring the attention of clinicians. This may increase the performance of clinicians to decide the condition of patient.

4.1 Limitations

The coverage of the ML-CDSS studies selected for this review may be incomplete. For example, we did not review conference abstracts or studies written in a language other than English. In addition, publication biases may occur because studies that report results showing that a ML-CDSS increased the performance are more likely to be published more frequently.

5 Conclusions

The aim of this review study was to analyse how experimental studies for measuring the performance of clinicians with or without the aid of ML-CDSS have been conducted. We explored key design factors and reviewed the choices made in previous studies regarding the factors. For example, dependent variables of experiment setup and available patient data determine performance measures that can be calculated. If the performance is measured by AUC values, probability/continuous input values (dependent variables) and retrospective patient data are required. In some studies more deep or subjective information, than just numerical values, were collected by mental efficiency type measures. The number of samples, number of participants and the duration of experiment are interrelated. The number of samples should be high enough to produce statistically significant results. However, increasing the number of samples per participant and thereby increasing the duration of experiment, can cause fatigue, which can lower the quality of input values. In many studies no patient background information was available for the participants. Such experiment setups differ from a real world patient examination and can bias the results.

Appendix: Search syntax for the Pubmed

("Machine Learning*"[Mesh] OR "Deep Learning"[MeSH] OR "Neural Networks, Computer*"[Mesh] OR scan assistant[Title/Abstract] OR “deep learning” [Title/Abstract] OR "random forest"[Title/Abstract] OR "support vector machine" [Title/Abstract] OR "decision tree"[Title/Abstract] OR "gradient boosting"[Title/Abstract])

AND

("Proof of Concept Study"[Mesh] OR "ROC Curve"[Mesh] OR “Prospective Studies” [Mesh] OR “with and without” [Title/Abstract] OR examination[Title/Abstract]) OR "controlled experiment"[Title/Abstract] OR “unblinded randomized clinical trial” [Title/Abstract] OR “prospective randomized controlled study” [Title/Abstract])

AND

("Decision Support Systems, Clinical"[Mesh] OR "Image Interpretation, Computer-Assisted"[Mesh] OR "Image Interpretation, Computer-Assisted/statistics and numerical data"[MAJR] OR "Image Processing, Computer-Assisted/methods"[MAJR] OR "clinical understanding" [Title/Abstract] OR “computer-aided diagnosis” [Title/Abstract] OR “early warning system” [Title/Abstract])

NOT

(Comment[Publication Type] OR editorial[Publication Type] OR letter[Publication Type] OR case reports[Publication Type])

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12553-023-00763-1.

Author contributions Mikko Nuutinen: Conceptualization, Methodology, Investigation, Writing - Original Draft. Riikka-Leena Leskelä: Writing - Review & Editing, Supervision.

Funding Open Access funding provided by University of Helsinki including Helsinki University Central Hospital. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreements No 777167 and No 965341.

Availability of data and material Not applicable.

Code availability Not applicable.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Conflict of interests The authors have no relevant financial or non-financial interests to disclose.

References

1. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci. 2018;115(45):11591–6.
2. Zhou QQ, Wang J, Tang W, Hu ZC, Xia ZY, Li XS, et al. Automatic detection and classification of rib fractures on thoracic CT using convolutional neural network: accuracy and feasibility. Korean J Radiol. 2020;21(7):869.
3. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut. 2019;68(12):2161–9.
4. Lundberg SM, Nair B, Vavilala MS, Horibe M, Eisses MJ, Adams T, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. Nat Biomed Eng. 2018;2(10):749–60.
5. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery: the HYPE randomized clinical trial. Jama. 2020;323(11):1052–60.
6. Watkinson P, Clifton D, Collins G, McCulloch P, Morgan L, Group DAS, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med. 2021;2021.
7. Ginestra JC, Giannini HM, Schweickert WD, Meadows L, Lynch MJ, Pavan K, et al. Clinician perception of a machine learning-based early warning system designed to predict severe sepsis and septic shock. Crit Care Med. 2019;47(11):1477.
8. Shen J, Zhang CJ, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: systematic review. JMIR Med Inform. 2019;7(3):e10010.
9. Groot OQ, Bongers ME, Ogink PT, Senders JT, Karhade AV, Bramer JA, et al. Does artificial intelligence outperform natural intelligence in interpreting musculoskeletal radiological studies? A systematic review. Clinical Orthopaedics and Related Research®. 2020;478(12):2751–2764.
10. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. BMJ. 2020;368.
11. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al. Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery. 2018;83(2):181–92.
12. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. BMJ. 2020;370.
13. 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. BMJ. 2020;370.
14. Brinker TJ, Hekler A, Hauschild A, Berking C, Schilling B, Enk AH, et al. Comparing artificial intelligence algorithms to 157 German dermatologists: the melanoma classification benchmark. Eur J Cancer. 2019;111:30–7.
15. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. Eur J Cancer. 2019;111:148–54.
16. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24(9):1342–50.
17. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, etal. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–118.
18. Han SS, Park GH, Lim W, Kim MS, Na JI, Park I, et al. Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: Automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PloS one. 2018;13(1):e0191493.
19. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine. 2019;9:52–9.
20. Lehman CD, Wellman RD, Buist DS, Kerlikowske K, Tosteson AN, Miglioretti DL, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med. 2015;175(11):1828–37.
21. Brocklehurst P, Field D, Greene K, Juszczak E, Keith R, Kenyon S, et al. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. Lancet. 2017;389(10080):1719–29.
22. Dhombres F, Maurice P, Guilbaud L, Franchinard L, Dias B, Charlet J, et al. A novel intelligent scan assistant system for early pregnancy diagnosis by ultrasound: clinical decision support system evaluation study. J Med Internet Res. 2019;21(7):e14286.
23. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol. 2018;42(12):1636.
24. Kiani A, Uyumazturk B, Rajpurkar P, Wang A, Gao R, Jones E, et al. Impact of a deep learning assistant on the histopathologic classification of liver cancer. NPJ Digital Medicine. 2020;3(1):1–8.
25. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(10):1813–9.
26. Bien N, Rajpurkar P, Ball RL, Irvin J, Park A, Jones E, et al. Deep-learning-assisted diagnosis for knee magnetic resonance imaging:

development and retrospective validation of MRNet. PLoS medicine. 2018;15(11):e1002699.

27. Su JR, Li Z, Shao XJ, Ji CR, Ji R, Zhou RC, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc. 2020;91(2):415–24.
28. Tajmir SH, Lee H, Shailam R, Gale HI, Nguyen JC, Westra SJ, et al. Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. Skelet Radiol. 2019;48(2):275–83.
29. Sim Y, Chung MJ, Kotter E, Yune S, Kim M, Do S, et al. Deep convolutional neural network-based software improves radiologist detection of malignant lung nodules on chest radiographs. Radiology. 2020;294(1):199–209.
30. Lee JH, Ha EJ, Kim D, Jung YJ, Heo S, Jang YH, et al. Application of deep learning to the diagnosis of cervical lymph node metastasis from thyroid cancer with CT: external validation and clinical utility for resident training. Eur Radiol. 2020;3066–3072.
31. Kozuka T, Matsukubo Y, Kadoba T, Oda T, Suzuki A, Hyodo T, et al. Efficiency of a computer-aided diagnosis (CAD) system with deep learning in detection of pulmonary nodules on 1-mm-thick images of computed tomography. Jpn J Radiol. 2020;38(11):1052–61.
32. Jang S, Song H, Shin YJ, Kim J, Kim J, Lee KW, et al. Deep learning-based automatic detection algorithm for reducing overlooked lung cancers on chest radiographs. Radiology. 2020;296(3):652–61.
33. Cha KH, Hadjiiski LM, Cohan RH, Chan HP, Caoili EM, Davenport MS, et al. Diagnostic accuracy of CT for prediction of bladder cancer treatment response with and without computerized decision support. Acad Radiol. 2019;26(9):1137–45.
34. Cai SL, Li B, Tan WM, Niu XJ, Yu HH, Yao LQ, et al. Using a deep learning system in endoscopy for screening of early esophageal squamous cell carcinoma (with video). Gastrointest Endosc. 2019;90(5):745–53.
35. Sato Y, Takegami Y, Asamoto T, Ono Y, Hidetoshi T, Goto R, et al. Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. BMC Musculoskelet Disord. 2021;22(1):1–10.
36. Yu Q, Huang K, Zhu Y, Chen X, Meng W. Preliminary results of computer-aided diagnosis for magnetic resonance imaging of solid breast lesions. Breast Cancer Res Treat. 2019;177(2):419–26.
37. Choi SY, Park S, Kim M, Park J, Choi YR, Jin KN. Evaluation of a deep learning-based computer-aided detection algorithm on chest radiographs: Case–control study. Medicine. 2021;100(16).
38. Choi JW, Cho YJ, Ha JY, Lee YY, Koh SY, Seo JY, et al. Deep learning-assisted diagnosis of pediatric skull fractures on plain radiographs. Korean J Radiol. 2022;23(3):343.
39. Shang S, Huang C, Yan W, Chen R, Cao J, Zhang Y, et al. Performance of a computer aided diagnosis system for SARS-CoV-2 pneumonia based on ultrasound images. Eur J Radiol. 2022;146:110066.
40. Roller R, Mayrdorfer M, Duettmann W, Naik MG, Schmidt D, Halleck F, et al. Evaluation of a clinical decision support system for detection of patients at risk after kidney transplantation. Front Public Health. 2022;10:979448. https://doi.org/10.3389/fpubh.2022.979448.
41. Wang X, Sun Z, Xue H, Qu T, Cheng S, Li J, et al. A deep learning algorithm to improve readers' interpretation and speed of pancreatic cystic lesions on dual-phase enhanced CT. Abdominal Radiology. 2022;47(6):2135–47.
42. Yacoub B, Varga-Szemes A, Schoepf UJ, Kabakus IM, Baruah D, Burt JR, et al. Impact of artificial intelligence assistance on chest CT interpretation times: a prospective randomized study. Am J Roentgenol. 2022;219(5):743–51.
43. Wei Q, Zeng SE, Wang LP, Yan YJ, Wang T, Xu JW, et al. The added value of a computer-aided diagnosis system in differential diagnosis of breast lesions by radiologists with different experience. J Ultrasound Med. 2022;41(6):1355–63.
44. Wataya T, Yanagawa M, Tsubamoto M, Sato T, Nishigaki D, Kita K, et al. Radiologists with and without deep learning-based computer-aided diagnosis: comparison of performance and interobserver agreement for characterizing and diagnosing pulmonary nodules/masses. Eur Radiol. 2023;33(1):348–59.
45. Toda N, Hashimoto M, Iwabuchi Y, Nagasaka M, Takeshita R, Yamada M, et al. Validation of deep learning-based computer-aided detection software use for interpretation of pulmonary abnormalities on chest radiographs and examination of factors that influence readers' performance and final diagnosis. Jpn J Radiol. 2022;1–7.
46. Chidambaram AG, Josephson M. Clinical research study designs: The essentials. Pediatric Investigation. 2019;3(4):245–52.
47. Jhangiani RS, Chiang ICA, Cuttler C, Leighton DC, et al. Research methods in psychology. Kwantlen Polytechnic University 2019.
48. Weerts HJ, van Ipenburg W, Pechenizkiy M. A human-grounded evaluation of shap for alert processing. arXiv preprint arXiv:190703324. 2019.
49. BT RIR. Methodology for the subjective assessment of the quality of television pictures. International Telecommunication Union. 2002.
50. Recommendation I. General methods for the subjective assessment of sound quality. ITU-R BS. 2003;1284–1.
51. Shafinah K, Selamat M, Abdullah R, Muhamad A, Noor A. System evaluation for a decision support system. Inf Technol J. 2010;9(5):889–98.
52. Yamashita K, Yoshiura T, Arimura H, Mihara F, Noguchi T, Hiwatashi A, et al. Performance evaluation of radiologists with artificial neural network for differential diagnosis of intra-axial cerebral tumors on MR images. Am J Neuroradiol. 2008;29(6):1153–8.
53. Pantanowitz L, Sinard JH, Henricks WH, Fatheree LA, Carter AB, Contis L, et al. Validating whole slide imaging for diagnostic purposes in pathology: guideline from the College of American Pathologists Pathology and Laboratory Quality Center. Arch Pathol Lab Med. 2013;137(12):1710–22.
54. Nielsen PS, Lindebjerg J, Rasmussen J, Starklint H, Waldstrøm M, Nielsen B. Virtual microscopy: an evaluation of its validity and diagnostic performance in routine histologic diagnosis of skin tumors. Hum Pathol. 2010;41(12):1770–6.
55. Schaaf J, Sedlmayr M, Sedlmayr B, Prokosch HU, Storf H. Evaluation of a clinical decision support system for rare diseases: a qualitative study. BMC Med Inform Decis Mak. 2021;21(1):1–11.
56. Das D, Chernova S. Leveraging rationales to improve human task performance. In: Proceedings of the 25th International Conference on Intelligent User Interfaces. 2020;510–518.
57. Lai V, Tan C. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency. 2019;29–38.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.