Effects of computer-augmented reporting on diagnostic accuracy and time consumption in thoracic radiology – a systematic review

Louis L. Plesner MD[1,2,3], Mia D. Jørgensen[1], Felix C. Müller MD, PhD [1,3], Mikael Boesen MD, PhD, Professor[2,3,5], Olav Wendelboe Nielsen MD, PhD, Professor[2,4], Michael B. Andersen MD[1,3]

[1] Department of Radiology, Herlev and Gentofte Hospital, Copenhagen, Denmark

[2] Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark

[3]Radiological AI testcenter, RAIT.dk, Capital region of Denmark

[4] Department of Cardiology, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark

[5] Department of Radiology, Bispebjerg and Frederiksberg Hospital, Copenhagen, Denmark

Corresponding Author:

Louis Lind Plesner, MD, Radiology Resident

Herlev Hospital, Copenhagen, Denmark

Telephone: 0045 3868 1517

Mail: louis.lind.plesner@regionh.dk // louislindplesner@gmail.com

Abstract

Introduction:

In recent years, there have been a tremendous growth of research on artificial intelligence (AI) in radiology. However, the vast majority of current AI products for detection of pathology, are intended to be used as an assisting tool for the radiologist and not as a stand-alone solution. The primary question of this systematic review was: Does computer-assisted reading (CAR) (deep-learning or other traditional CAD methods) affect diagnostic accuracy and or reading time versus normal reading in daily radiological workflow in thoracic radiology.

Methods:

The study is a systematic review following the PRISMA statement. We shall select publications based on following inclusion criteria: English language; peer reviewed scientific diagnostic accuracy study, quality improvement study or RCT; assessment of the difference in diagnostic accuracy and/or time consumption of an AI or traditional CAD algorithm assisting a physician for diagnosis in thoracic/chest radiological imaging versus standard reporting. Primary outcome is the difference in sensitivity and specificity of CAR versus unassisted reading in different target conditions and imaging methods. Secondary outcomes are time consumption and risk of bias in included studies (using QUADAS-2). Studies will be compared using descriptive statistics on study design parameters and one or more metaanalyses, if the studies are comparable on key parameters.

Conclusion:

This systematic review will assess the effects of computer-assisted reading using AI products in thoracic radiology on diagnostic accuracy and establish grounds for further research in this field.

Introduction

In recent years, there have been a tremendous growth in research on artificial intelligence (AI) for various health applications. Several specialties including radiology, dermatology and ophtalmology have gained particular attention due to the increasing performance of especially deep-learning (DL) algorithms in the field of computer vision, a particular subset of AI dealing with image analysis[1–3]. These technologies has led to studies showing increasing diagnostic accuracy by algorithms for image-based classifications tasks for multiple purposes in health research[1,3]. However, difficulties in external generalizability and stability over time combined with limited translational research and research of implementation workflow likely has contributed to the lack of widespread adoption of these techniques[4]. Contributing to this, there have been published a very limited number of systematic reviews on AI performance in radiology, though several recent guidelines have established advise on data management, ethics, study designs and potential future research[2,5–7].

Although AI has shown promise in several specific domains, there is likely still a long way before AI could function independently or minimally dependently of human interaction[2,3]. Therefore, the vast majority of current AI products for detection and classification of pathology in radiological studies, are intended to be used as an assisting tool for the radiologist[8]. Using this workflow, it is the radiologist who carries responsibility for image interpretation including dismissal of potential false positive and false negative markings by the algorithm. When searching through the list of FDA classified AI algorithms for chest imaging (03 May 2021), currently 28 AI products can be found, of which (from 'Indications For Use') not a single product could function independently without co-reading of the study by the radiologist. The claim by the manufacturer is in most cases, that diagnostic accuracy can be approved when using these algorithms for 'double reading'. However,

this needs to be measured against potentially increased reading time and lowered specificity due to possible trust in false positives by the algorithm. Furthermore, false positive markings may exert fatigue on the radiologists reading the study and thereby lowering work satisfaction. Detection and classification of radiological findings using computer-assisted methods is not new. Computer-Assisted Diagnosis (CAD) software has existed for decades, but has seen vastly further promise given the more powerful modern DL technology, such as convolutional neural networks (CNN).

The primary research question if this systematic review was: Does computer-assisted reading (using deep-learning or other traditional CAD methods) affect diagnostic accuracy and or reading time versus normal unassisted reading in daily radiological workflow in thoracic radiology; and further, in which use-cases and workflows might this be the case? Implied important questions: do these algorithms increase sensitivity and how are physicians (radiologists and non-radiologists) affected by false positive markings? Additionally, based on systematic characterization the quality of evidence, assessing study design and risk of bias in the studies (using the TRIPOD guideline as reference), we aim to recommend directions for further research in this area.

Methods

Study design

The study is a systematic review following the PRISMA statement. Furthermore, we shall perform one or more metaanalyses, if the studies in the literature are found comparable. Specifically, a metaanalysis will be performed if there is found more than one study for a particular modality (ie. computed tomography (CT) or radiography); type of algorithm (ie. CNN or traditional CAD) and

intended use (ie. pneumothorax, COVID-19, pulmonary embolism or nodule) using a consecutive external patient cohort with comparable reader experience (ie. senior radiologists or residents). If there is not sufficient evidence to perform a meta-analysis, a narrative synthesis of the systematically collected data shall be performed.

Study identification and inclusion criteria

We will perform a comprehensive search by using broad keywords to identify eligible studies. The following electronic databases will be searched from 2010 to March 2021 (and updated before submission): Medline, Embase, SCOPUS and Web of Science. Additional articles will be retrieved by manually going through the reference lists of included publications. We shall select publications for review if they satisfy several inclusion criteria: English language; peer reviewed scientific diagnostic accuracy study, quality improvement study or RCT; assessment of the difference in diagnostic accuracy and/or time consumption of an ML or traditional CAD algorithm assisting a physician for diagnosis in thoracic/chest radiological imaging versus standard reporting. Both arms should be compared to a reference standard ('ground truth') by at least one human considered an expert in the field. Furthermore, CT, biopsy results or – in case of COVID-19 studies – RT PCR, should be included (where appropriate) in reference standard labelling. Exclusion criteria includes non peer-reviewed publication types (commentaries, letters, conference papers, meeting abstracts); not satisfactory reference standard; no provided experience level of readers, not available sensitivity or specificity measurements and studies using computer generated non-trivial post-processing techniques to modify images.

Computer-assisted reporting (CAR) was defined as reading of radiological studies using an autogenerated report and/or image overlay (ie. heatmap) by a machine learning algorithm, including deep-learning (ie. CNN and others) or traditional computer-aided-diagnosis (CAD) to broaden the field of included studies. Machine learning for the purpose of medical imaging can be defined as models composed of multiple processing layers to learn representations of data with multiple levels of abstraction without being explicitly programmed. CAD was defined as a computer generated output (by other methods than neural networks) to assist the clinician in making a diagnosis. For all types of algorithms studies of computed-aided detection (where a lesion is flagged, but no diagnosis suggested) and computer-aided diagnosis (where a lesion is classified in pathology category) where considered eligible. The minimum requirements for measurements of diagnostic accuracy in included studies are sensitivity, specificity, in total numbers or divided in subgroups. For example, studies reporting only sensitivity and false positives per image, without providing data for possible calculation of specificity will be excluded. In studies from 2017 and forward, corresponding author will be contacted with request of further data before such exclusion.

Study selection and extraction of data

Two researcers (LLP and MDJ) will independently screen abstracts for potentially eligible studies. In the identified studies, full text reports shall be assessed by the same researchers for eligibility with disagreements resolved by consensus. Furthermore, the same two researchers will extract data from study reports independently, with disagreements resolved by consensus or by consulting a third reviewer (MBA).

Adherence to reporting standards and risk of bias

All included studies shall be assessed for reporting standards using a modified version of the TRIPOD checklist. For the purpose of this systematic review, not all items were considered relevant, in particular should be mentioned that we do not asses the items regarding algorithm development, training, neural network structure etc. This is because the objective was not to examine the algorithm performance in itself, but rather the effect of the algorithm on the clinician. Potential bias in algorithm development should therefore only weaken algorithm performance, if the reader test is performed in an external validation cohort. Furthermore, we shall perform a risk of bias assessment of each study using an adapted QUADAS-2 tool (Figure 1).

Outcomes

Primary: The difference in sensitivity, specificity, AUC of computer-assisted human reader versus human reader alone.

Secondary: Time consumption with and without computer-assisted reading

Qualitative/methodological outcomes: Risk of bias (QUADAS-2 in studies), Study design types, ie. adherence to items in table 1 (modified from TRIPOD guideline)

Statistics

The primary outcome is the difference sensitivity, specificity and/or AUROC curve (area under receiver operating characteristics curve) when reading studies unaided versus computer assisted. If

studies meet the criteria mentioned in 'study design', they will be included in a metaanalysis. If data allows, we are planning to do subgroup analysis of the effect of AI in radiology residents vs. radiology specialists vs. non-radiologist physicians and consecutive versus enriched cohort. However, if there are not found sufficient studies to perform one (or more) metaanalysis, the results will be displayed in table format with primary and secondary outcomes. Study design characteristics will be reported using percentages (categorical variables) or mean (SD)/median (IQR) (continuous variables). For any statistical analysis shall use Rstudio (v. 3.6.1).

Conclusion

We have designed a systematic review to assess the current literature for studies describing the potential of computer-assisted reading by radiologists. Furthermore, by assessing the potential biases in such studies we should be able to recommend future directions for research of computer-assisted reading by radiologists, in the era of deep-learning.

Table 1: Items for dataextraction in review.

Introduction
Target condition
Modality
Primary outcome
Country of last author
Methods
Prospective/retrospective/rct/other(specify)
Multicenter/single center
Setting (ED,outpatients,mixed)
Diagnostic, control of procedure, control
Consecutive, enriched, case-control, convenience sample of patients
Concurrent or second reader
Research or clinical setting
Time restriction yes/no
No. of readers, training and experience
Access to other patient data, file, priors etc?
Ground truth reference standard
Sample size calculations yes/no
External, in-house, internal split dataset validation or combination
Type of model - eg. CNN or CAD
Name of algorithm incl. version
Results
No. of included patients (total)
No. of included patients (reader performance subset)
Clinical information provided on patients (age,

sex, other)
Performance human reader(s) alone (eg. n, AUC, sens, spec, PPV, NPV)
Performance human + AI (eg. n, AUC, sens, spec, PPV, NPV)
Difference AUC etc. incl. P values (S or NS)
Time difference incl. p value (S or NS) Number of FP pr. 100 cases (calculated from spec.)

QUADAS-2 assessment

Research question: Does computer-assisted reading affect diagnostic accuracy or reading times in daily radiological practice?

Domain 1 – Patient Selection (HIGH/LOW/UNCLEAR)

Signaling question 1: Was a consecutive, random or enriched sample of patients enrolled? (HIGH/LOW/UNCLEAR)

For the purpose of this review, consecutive and random types of patient cohorts will be accepted as without risk of bias, given the effect is the change in diagnostic accuracy, and not diagnostic accuracy in absolute numbers. Enriched samples will have high risk of bias, because it likely does not reflect true working conditions.

Signaling question 2: Was a case-control design avoided? (HIGH/LOW/UNCLEAR)

Case-control design: High risk of bias

Signaling question 3: Did the study avoid inappropriate exclusions? (HIGH/LOW/UNCLEAR)

Signalling question 4: Were the same patients included several times in pooled analysis (eg. 1 time for each reader)? (HIGH/LOW/UNCLEAR)

If the study includes multiple readers, it should be noted if they read the same cases or different cases from each other. If they read the same cases, the cases should not count several times in study n → high risk of bias.

Applicability: Are there concerns that the included patients and setting does not match the review question? (HIGH/LOW/UNCLEAR)

Domain 2 – Index test (HIGH/LOW/UNCLEAR)

Signaling question 1: Were the index test results interpreted without knowledge of the results of the reference standard? (HIGH/LOW/UNCLEAR)

Index test here is the diagnostic accuracy/reading time with/without computer-assisting.

Signalling question 2:

Was there a reasonable time frame between reporting with and without AI? (HIGH/LOW/UNCLEAR)

This should be at least 4 weeks.

Signaling question 3: If a threshold was used, was it prespecified? (HIGH/LOW/UNCLEAR)

AI studies often report on high and low sensitivity thresholds, they should be prespecified.

Signalling question 4:

Was the algorithm for index test developed on same data as the test ie. internal validation? (HIGH/LOW/UNCLEAR)

Internal validation studies will have high risk of bias, due to overperformance of algorithm.

Applicability: Are there concerns that the index test, its conduct, or its misinterpretation differ from the review question? (HIGH/LOW/UNCLEAR)

Domain 3 – Reference standard (HIGH/LOW/UNCLEAR)

Signaling question 1: Is the reference standard likely to correctly classify the target condition? (HIGH/LOW/UNCLEAR)

For this review, the reference standard is considered majority vote and/or consensus review by expert radiologists and/or superior imaging method eg. CT for chest radiography. Reference standard label by only one expert radiologist should be judged with risk of bias. Furthermore, if ground truth diagnosis is not a radiological finding – but a specific disease, the ground truth should be made accordingly (e.g tuberculosis, COVID-19, lung carcinoma)

Signaling question 2: Were the reference standard results interpreted without knowledge of the index test? (HIGH/LOW/UNCLEAR)

Applicability: Are there concerns that the target condition as defined by the reference standard does not match the review question? (HIGH/LOW/UNCLEAR)

Domain 4 – Flow and timing (HIGH/LOW/UNCLEAR)

Signaling question 1: Was there an appropriate interval between the index test and the reference standard? (HIGH/LOW/UNCLEAR)

In retrospective imaging studies this will likely be low.

Signaling question 2: Was there an appropriate flow when reporting with and without AI? (HIGH/LOW/UNCLEAR)

Reading with AI should be done first in half of cases, or AI results may be exaggerated also due to the fact that readers have seen the scan before, and may notice new additions anyway. If these two conditions are not satisfied → high risk or unclear

Signaling question 3: Did all patients receive the same reference standard? (HIGH/LOW/UNCLEAR)

Signaling question 4: Were all patients included in the analysis? (HIGH/LOW/UNCLEAR)

Often in these studies, only a subset of cases goes through review with and without AI. This is fine, but it should be noted which cases are included in these analyses (random/enriched/consecutive).

Applicability: Are there concerns that the flow and timing in this study does not match the review question? (HIGH/LOW/UNCLEAR)

Appendix 2 TRIPOD items

Title
Abstract
Introduction - context
Introduction - objectives
Methods - study design
Methods - study dates
Methods - study setting
Methods - eligibility criteria
Methods - outcome predicted
Methods - sample size
Methods - missing data
Methods - model buidling
Methods - validation predictions
Methods - model performance
Results - flow of data
Results - characteristics
Results - numbers
Results - model performance
Discussion - limitations
Discussion - interpretation
Discussion - clinical use
Supplementary data
Funding
-----Flowchart?