Current clinical applications of AI in Radiology and their best supporting evidence

Amara Tariq, PhD[1], Saptarshi Purkayastha, PhD[2], Geetha Priya Padmanaban, MS[2], Elizabeth Krupinski, PhD[3], Hari Trivedi, MD[1,3], Imon Banerjee, PhD[1,3], Judy W.Gichoya, MBChB MS[1,3]

[1]Department of Biomedical Informatics, Emory School of Medicine, Atlanta, USA; [2]School of Informatics Computing, Indiana University Purdue University, Indianapolis, USA [3]Department of Radiology, Emory School of Medicine, Atlanta, USA

Abstract

Purpose: *Despite tremendous gains from deep learning and the promise of AI in medicine to improve diagnosis and save costs, there exists a large translational gap to implement and use AI products in real-world clinical situations. Adoption of standards like the TRIPOD, CONSORT and CLAIM checklists is increasing to improve the peer review process and reporting of AI tools. However, no such standards exist for product level review.*

Methods: *A review of the clinical trials shows a paucity of evidence for radiology AI products; thus, we developed a 10-question assessment tool for reviewing AI products with an emphasis on their validation and result dissemination. We applied the assessment tool to commercial and open source algorithms used for diagnosis to extract evidence on the clinical utility of the tools.*

Results: *We find that there is limited technical information on methodologies for FDA approved algorithms compared to open source products, likely due to concerns of intellectual property. Furthermore, we find that FDA approved products use much smaller datasets compared to open source AI tools, as the terms of use of public datasets are limited to academic and non-commercial entities which precludes their use in commercial products.*

Conclusion: *Overall, we observe a broad spectrum of maturity and clinical use of AI products, but a large gap exists in exploring actual performance of AI tools in clinical practice.*

Introduction

Clinical validation of artificial intelligence (AI) systems involves evaluation of their performance to meet a clinical need, through "systematic and planned processes to continuously generate, analyze and assess clinical data to verify the safety and performance [of the system], including clinical benefits"[1]. Between 2000 and 2018, there were 8813 radiology AI publications worldwide, 16.5 % of which were from the U.S.A [2]. Despite many publications on radiology AI, as of June 2020 there are only 62 U.S. Food and Drug Administration (FDA) approved AI applications for clinical usage [3], reflective of the challenge in obtaining regulatory approval for AI products. Even after this initial step of FDA approval, there remains a translational gap to enable actual use of the system in clinical practice which includes post market surveillance, software updates and adjustments to account for shifts in technical parameters or patient populations. In fact, a review of 516 studies published between January 2018 and August 2018 found that only 6 % (n = 31) of studies reported external validation with multi-institutional data or prospective validation. [4].

AI systems which perform well on the internal dataset used for validation, may not generalize well to new data, as demonstrated by a drop in performance when deployed into clinical workflow [5, 6]. Clinical standards and guidelines continuously change over time; including changes in treatment pattern, coding systems (with shift from ICD-9 to ICD-10), implementation of new medical records systems, new imaging equipment and protocols, or change in the incidence and prevalence of disease. While humans are adaptable to these changes, AI systems may falter because FDA-approved models cannot be significantly adjusted without losing certification. In addition, bias is noted as a problem in many AI systems, and this may not be identifiable during model training. An analysis of a commercial risk prediction tool used on 200 million people in the U.S. for high risk patient management shows significant racial bias, with black patients sicker than white patients at any given risk score [7]. This bias arose from predicting health costs rather than illness resulting in unequal access, and changing the outcome metric used for prediction increased the percentage of black patients receiving additional help from 17.7 % to 46.5 % [7].

Overall, there is a critical need for comprehensive review of AI tools beyond statistical validity of models (usually receiver operating characteristics (ROC) curves, specificity, sensitivity, accuracy, and positive and negative predictive values), to include clinical validation that evaluates the model performance when deployed to actual clinical settings. Such tools are generally not available for use, however there are several efforts being made to improve standardized reporting of AI including adoption of Standards for Reporting of Diagnostic Accuracy Studies (STARD) [8], Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) [9] and Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [10]. However, these tools are geared towards peer reviewed articles and clinical trials reporting, and are not applied to commercial products. We expand upon these reporting tools to develop a new questionnaire for product evaluation that can be applied to directly to FDA-cleared products rather than manuscripts or other peer-reviewed activity. We apply this questionnaire to FDA-approved algorithms through April 2020 as well as on open source systems AI tools for which sufficient methodological data is available.

Methodology

For this paper, we focus on commercial and open source AI tools applied directly to images to facilitate diagnosis. There are a wide variety of proprietary and open-access tools focused on interpretation of radiology images using classification algorithms. In this section, we present the details of how we gathered information for FDA-cleared AI algorithms as well as open-source AI tools for diagnosis. Note that we concluded our search in April 2020. Therefore, tools developed or published after this date are not part of our review. Patients were not involved in the study design, conduct or evaluation, thus the review did not require IRB approval.

The STARD2015, CLAIM and TRIPOD checklists used for reporting of results are very detailed and cover many areas including *Methodology*, *Study Design*, *Participants*, *Test Methods*, and *Results*. We prepared a 10-question, open-ended assessment instrument (see Appendix A in supplementary material) that combines important criteria from the STARD2015 [8], TRIPOD [9] and CLAIM [10] checklists to perform a comprehensive review of the AI tools. Because many of the tools we assessed are commercial and hence proprietary, we realized that many details about their algorithms could not be shared due to company policy. Therefore, for this questionnaire/assessment we focused on the *Results* section of these guidelines and asked for minimal information regarding technical details of the algorithm. We also include a section regarding *Dissemination* that includes questions about publications, public datasets and participation in public AI challenges.

FDA-cleared AI Tools: FDA has been updating its policies to keep up with the dynamic nature of development and evaluation of software tools, termed Software as Medical Device (SaMD) [11]. Manufacturers are required to file marketing application (510(k) notification, De Novo, or premarket approval application (PMA) pathway) with FDA prior to distribution of their device. Type of submission and data requirements change based on the risk category of SaMD. Risk categorization described by International Medical Device Regulatory Forum (IMDRF) is based on intended medical purpose (treat, diagnose, drive clinical management, inform clinical management) and healthcare situation (critical, serious, non-serious) of SaMD. IMDRF also describes three major aspects of clinical evaluation of SaMD, i.e., clinical association (valid clinical association between SaMD output and targeted clinical condition), analytical validation (correct processing of input data to generate accurate, precise and reliable output data) and clinical validation (achievement of intended purpose in targeted population in the context of clinical care using SaMD output).

We used the list of FDA-approved AI tools maintained by Data Science Institute of American College of Radiology [1] which contained 45 tools from 32 different companies as of April 2020. Even though the list contains a wide variety of imaging types that target many body parts, we identified the following major areas of interest which have the highest number of commercial products.

- Computerized tomography (CT) and magnetic resonance imaging (MRI) of the head
- CT and X-ray of the chest
- Mammography and ultrasound of the breast

Our assessment instrument was shared with companies with FDA-approved AI tools for completion.

Open-source AI Tools: We searched PubMed for open-source AI algorithms used in radiological image analysis in the three main areas of interest identified above. We mainly focused on peer-reviewed publications with open-access to code and datasets. We included a few major publications that showed openness in terms of sharing this information through an *access upon-request* clause. For these tools, we completed the assessment ourselves based on the data provided in the publications, appendices, project pages and public code bases.

[1] https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms

Results

FDA approved tools

Table 1 shows a list of companies with FDA-approved tools who responded to our request for information. We were able to collect information on 45% of the FDA-approved tools mentioned in the list maintained by Data Science Institute of the American College of Radiology. A detailed description of their technical details in not possible due to concerns over Intellectual Property. A summary of the assessment responses for these tools are provided in Tables 2, 3 and 4.

Company	AI Tool
RADLogics	AI Medical Imaging (AIMI) Platform
Imaging Biometrics, LLC	IB Neuro IB Stone Checker
Koios Medical, Inc.	Koios DS for Breast
Quantib	Quantib Brain
iCAD Inc.	PowerLook Tomo Detection V2 Software
Subtle Medical	SubtleMR SubtlePET
Vital Images	Vital CT Lung Density Analysis Vital CT Brain Perfusion
Zebra Medical Vision	HealthCCS HealthCXR HealthICH HealthPNX HealthVCF
AIDOC	AIDOC-Briefcase-ICH AIDOC-Briefcase-CSF AIDOC-Briefcase-PE AIDOC-Briefcase-LVO

Table 1: List of FDA approved companies who completed the 10-question assessment instrument.

Open Sources Tools

Our selection of open source tools was based on the three anatomical areas covered by FDA approved tools - chest, head and breast. Table 5 provides a summary of the assessment conducted by our team on the open-source AI tools. Because the open-source tools have details on model architecture, we have included this information here whereas it is not present for commercial companies.

A summary of multiple high impact papers grouped by body region are provided below.

Chest

We reviewed three major open-access tools and peer-reviewed publication for chest imaging analysis. The first one is COVID-NET, a deep learning-based model for detection of COVID-19 through chest X-ray analysis. It was developed by researchers - Linda Wang and Alexander Wong from Vision and Image Processing Lab of the University of Waterloo, Canada [12]. They have made their code and data publicly available in GitHub [2]. The second tool reviewed is Chester-The Radiology Assistant [13], an open-access chest X-ray analysis software developed by research at Montreal Institute of Learning Algorithms (MILA) in Montreal, Canada. The third one is a recently published work describing the development of PENet, a deep learning model for detection of pulmonary embolism through chest CT analysis [14].

[2] https://github.com/lindawangg/COVID-Net

[3] https://github.com/nyukat/breast_cancer_classifier

Breast

We reviewed five open access tools and peer-reviewed publications for breast imaging. The first tool was developed by researchers at New York University (NYU) using deep learning for breast cancer detection on mammograms [15]. Their code and pre-trained models are publicly available[3]. The second tool was from Google DeepMind, who published a peer-reviewed article on generalization of deep learning-based breast cancer detection [16] using datasets from United States and United Kingdom. Their code is not publicly available, but the model is described in detail in their paper, allowing for replication/implementation. The third tool was by researchers from Houston Methodist, Texas and Far-Eastern Memorial Hospital, Taiwan, who developed a breast cancer risk evaluation model that combined imaging data, risk factors from clinical reports and patient's demographic information [17]. Authors have announced development of a public-access application upon completion of their research. The fourth tool was from Yala et al who published an early-detection model for breast cancer patients using a dataset with five-year follow-up information from patients [18]. Their project code is available [4]. The fifth tool reviewed was from Shen et al who published a peer-reviewed article describing their unique deep learning-based model that learns to generalize from patch-based lesion detection for whole image breast cancer screening [19]. Their code is publicly available [5].

Head

We reviewed two major publications that apply deep learning-based AI analysis of head CT scans for detection of intracranial hemorrhage (ICH). The first model is by Kuo et al. who trained a patchFCN model that processes CT scan patch-by-patch through a fully-connected network [20]. The article provides evidence that detection of ICH cannot be performed the same way as object detection. This is because typical object detection datasets contain solid objects with particular shapes, and hence they can be first detected and then localized in natural images. On the other hand, ICH appears as *fluid* objects with no particular shape. Thus, it is better to detect them in all patches of the scan. The second model reviewed is by Arbabshirani et al. who trained a three-dimensional convolutional neural network (CNN) for detection of ICH and integrated their model in the clinical workflow to re-prioritize studies based on detected abnormalities [21]. They collected approximately 46583 CT scan studies from multiple locations to develop their model. They report several cases where their AI model was able to re-prioritize *routine* studies to *stat* statusand help reduce diagnosis time significantly.

[4] http://learningtocure.csail.mit.edu/

[5] https://github.com/lishen/end2end-all-conv

AI Product	Company	Target Area	Modality	Application	Validation Dataset	Performance
IB Stone Checker	Imaging Biometrics, LLC	Abdomen	CT	Kidney stone physical attribute characterization such as volume, mean HU density, skin-to-stone distance, entropy,kurtosis, and skewness	Usage validation (evaluating and usability testing) at two clinical sites in Oxford, UK and Beijing, China	92% accuracy in predicting number of shocks needed to fragment the stone, superior to the use of stone volume or density
IB Neuro	Imaging Biometrics, LLC	Brain	MRI	Quantitative Perfusion MRI algorithm that accounts for creation of quantitative (standardized) rCBV maps independent of scanner field strength, platform, or time point.	0, 11, 9, and 23 patients with tumors classified as WHO grades I, II, III, and IV, respectively; 30 male, 13 female	96% for high grade and 69% for low grade tumor prediction accuracy, by using a combination of rCBV and a vessel size index derived from a combined GE/SE pulse sequence.
Koios DS for Breast	Koios Medical, Inc.	Breast	Ultrasound	Detection and characterization of breast cancer	900 lesions from 900 patients	AUC: 0.882
Quantib Brain	Quantib	Brain	MRI	Automatic labeling, visualization, and volumetric quantification of segmentable brain structures, Atrophy detection, WMH Quantification	For brain tissue: 33 MR images, For WMH analysis: 45 3D T1w images	Visual inspection of 209 segmentations by experts: No obvious errors in 98% brain tissue and 97% WML segmentations
PowerLook Tomo Detection V2 Software	iCAD Inc.	Breast	DBT	Analysis 4-view tomosynthesis cases	260 Hologic DBT cases	Clinical Performance benefits of using CAD 8.0% sensitivity increase, 6.9% specificity increase, 7.2% reduction in recalls, 5.7% improvement in AUC, 52.7% reduction in reading time
WMH: White Matter Hyperintensity, WML: White Matter Lesions DBT: Digital Breast Tomosynthesis, rCBV: Relative Cerebral Blood Volume						

Table 2: FDA-approved proprietary AI tools for Radiological Image Analysis - Part A

AI Product	Company	Target Area	Modality	Application	Validation Dataset	Performance
Vitrea CT Lung Density Analysis	Vital Images, Inc	Chest	CT	Analysis of lung densities and volumes by segmenting lung tissues (semi-automated)	14 bilateral LTx recipients (66 with CLAD and 48 Stable), 23 single LTx recipients (11 with CLAD, 12 Stable)	AUC = 0.89 for CLAD prediction in single LTx, AUC = 0.63 for CLAD prediction in bilateral LTx.
Vitrea CT Brain Perfusion	Vital Images, Inc.	Brain	CT	Post-processing calculation of CBF, CBV, local bolus timing, MTT. displays time density curves, perfusion and summary maps, regions of interest and mirrored regions	183 patients underwent multimodal stroke CT using a 320-slice scanner within 6 hours of acute stroke onset, followed by 24 hour MRI that included DWI and PWI	The probability model was accurate at detecting ischemic core (AUC = 0.80, SD = 0.75-0.83) and penumbra (AUC = 0.85, SD = 0.83-0.87) and was significantly closer in volume to the reference DWI (P=0.031).
HealthCCS	Zebra Medical Vision	Heart/Chest	CT	Post-processing software for calcified plaque in coronary arteries, Categorization into 4 risk categories	249 studies of patients aged 20 years and above	Device-expert agreement: 89%
HealthCXR	Zebra Medical Vision	Chest	X-ray	Assessment of features indicating PEF	554 chest X-ray, Groundtruth established by experts	Operating Point#1: Sensitivity=96.74% Specificity=93.17% Operating Point#2: Sensitivity=93.84% Specificity=97.12%
HealthICH	Zebra Medical Vision	Head	CT	Assessment of features indicating ICH	427 head CTs, Groundtruth established by experts	Sensitivity= 95.11% Specificity= 91.98%
HealthPNX	Zebra Medical Vision	Chest	X-ray	Assessment of features indicating PNX	588 chest X-rays, Groundtruth established by experts	Sensitivity=93.15% Specificity=92.99%
HealthVCF	Zebra Medical Vision	Chest Abdominal	CT	Assessment of vertebral compression fracture	611 chest/abdominal CT, Groundtruth established by experts	Sensitivity=90.20% Specificity=86.89%

CLAD: Chronic Lung Allograft Dysfunction, LTx: Long-term Survival after Lung Transplantation
CBF: Cerebral Blood Flow, CBV: Cerebral Blood Volume, Local Bolus Timing: Delay of tissue response, MTT: Mean Transit Time
DWI: Diffusion Weighted Imaging, PWI: Dynamic Susceptibility Weighted Perfusion Imaging
PEF: Pleural Effusion, ICH: Intracranial Hemorrhage, PNX: Pneumothorax

Table 3: FDA-approved proprietary AI tools for Radiological Image Analysis - Part B

AI Product	Company	Target Area	Modality	Application	Validation Dataset	Performance
AlphaPoint Imaging Software	RAD-Logics	Chest	CT	Detection and quantification of lung nodules, PTX, enlarged heart	**PTX**: 300 cases. 158 positive, 142 negative, 168 male (Age: Mean = 51.6, SD=18.6, range= 18-91), 132 female (Age: Mean = 51.8, SD=16.2, range 23-86) **Corona**: 109 COVID-19 Chinese patients, 90 Patients with fever and upper respiratory tract symptoms, 49 patients classified by a radiologist as severe (n=13) vs non-severe (n=36)	**PTX**: AUC = 0.967 **Corona**: AUC = 0.948
SubtleMR	Subtle Medical, Inc.	Head, Spine, Neck, Knee	MRI	Enhance the MRI images by reducing image noise or by increasing image sharpness for non-contrast enhanced head MRI	11 consecutive patients (Age: 48+/-15 years; 7 female) undergoing clinical brain 1.5T MRI exams underwent an accelerated 3D sagittal FLAIR scan (average scan time reduction 27.1% +/-3.5%)	CNN based deep learning image processing of 3D FLAIR brain MRI. A boost in perceived image quality, SNR, and resolution despite a 30% reduction in scan time.
SubtlePET	Subtle Medical, Inc.	All body areas	FDG, PET, PET/CT, PET/MR	Quality enhancement of 2-fold, 3-fold, and 4-fold accelerated whole-body PET acquisitions	7 subjects (5 males, 2 female) referred for a whole-body FDG-18 PET/CT scan on a GE Discovery 710 scanner	All deep learning enhanced images (2 to 4-fold) demonstrated similar perceptual image quality and lesion conspicuity when compared to standard of care scans.
AIDOC Briefcase-ICH	AIDOC	Head	CT	Prioritization and flagging of ICH	7112 non-contrast head CT from two centers	Sensitivity: 95%, Specificity: 99%
AIDOC Briefcase-CSF	AIDOC	Cervical Spine	CT	Prioritization and flagging of cervical spine fractures	186 cases from two US sites and one site outside of US	Sensitivity: 91.7%, Specificity: 88.6%
AIDOC Briefcase-PE	AIDOC	Lungs	CTPA	Prioritization and flagging of pulmonary embolism	2915 CTPA, groundtruth established by experts	Sensitivity: 93%, Specificity: 95%
AIDOC Briefcase-LVO	AIDOC	Head	CTA	Prioritization and flagging of large vessel occlusion	338 cases from three US-based sites	Sensitivity: 88.8%, Specificity: 87.2%
CTPA: CT Pulmonary Angiogram						

Table 4: FDA-approved proprietary AI tools for Radiological Image Analysis - Part C

AI Product	Target Area	Modality	Model Type	Validation Dataset	Performance
Chester [13]	Lungs	Chest X-ray	DenseNet: disease prediction	ChestX-ray14 [22][public] PadChest [23][public]	AUC: 0.72-0.93 for different disease labels
COVIDNet [12]	Lungs	Chest X-ray	Deep CNN: Covid-19 detection	COVIDx* [public]	Sensitivity: 96.8%
PENet [14]	Lungs	CT Pulmonary Angiography (CTPA)	3D-CNN: PE detection	CTPA collected from two institutes [can be requested]	AUC: 0.84
NYU Breast Cancer Screening [15]	Breast	Mammogram	ResNet: Cancer Detection	229, 426 Mammography Studies	AUC: 0.895
BRISK Model [17]	Breast	Mammogram, Patient Records	Auto-Encoder	5174 patient records from Houston Methodist	AUC: 0.93 Accuracy: 81%
MIT Breast Cancer High Risk/Early Detection [18]	Breast	Consecutive Mammographic Studies	HybridDL: Early Detection using imaging and traditional risk factors	Consecutive screening of 39571 patients	AUC:0.70
DeepMind: International Breast Cancer Screening [16]	Breast	Mammogram	MobileNet ResNet	US dataset UK dataset**	AUC:0.87
End-to-end Approach for breast cancer screening [17]	Breast	Mammogram	CNN	CBIS-DDSM [24]	AUC: 0.98
DL-based ICH detection [20]	Brain	Head CT	PatchFCN	UCSF - 4400 CTs [can be requested]	AUC:0.99
ICH detection and workflow integration [21]	Brain	Head CT	3D-CNN	46583 CT scans collected from multiple facilities [can be requested]	AUC:0.85

*https://github.com/lindawangg/COVID-Net/blob/master/docs/COVIDx.md

**https://medphys.royalsurrey.nhs.uk/omidb/getting-access/

Table 5: Open-access AI tools for Radiological Image Analysis

Discussion

The gold standard for evidence in medicine is clinical trials. A search of the PubMed database reveals that although there many articles that report use of AI in radiology (n=2,067; keywords=”Radiology” + ”AI”), only about 2% discuss clinical practice of AI (n=40; keywords=”Radiology” + ”AI” + ”clinical practice” OR ”clinical trials”). More so, closer examination of these papers reveal that most are opinions and not reports of actual implementation. In the clinicaltrials.gov database (https://clinicaltrials.gov/ct2/results?term=radiology+AND+AI), 66 trials are registered, of which 10 are related to AI in clinical practice, but all of these trials were *recruiting*, *not yet recruiting* or *active but not recruiting*. Thus, we did not have enough data from these to identify clinical relevance of these AI tools.

Therefore, to obtain the next level of evidence, we adopted standardized checklists used for clinical trial reporting and peer review articles for meta-analysis, and developed a 10-question assessment tool to evaluate AI products. The broad themes of the assessment tool include *model type, dataset size and distribution, dataset demographics/subgroups, standalone model performance, comparative performance against a gold standard, failure analysis, publications, participation in public challenges, dataset release and scale of implementation*.

From the FDA tools we analyzed, we did not acquire technical details on the model type due to IP concerns. Open source tools made their code and model weights public through GitHub repositories (as mentioned in Table 2). Some of the open source tools were based on limited-access datasets collected from collaborating healthcare providing facilities. In these cases, even if the data were not made public, the model architecture was discussed in-depth, such that they are reproducible.

In response to our request, some companies provided details of the data they used for technical evaluation, as has been listed in Tables 3, 4, and 5. However, information on training datasets was not provided. Overall, the FDA approved tools used smaller datasets for validation compared to open source tools whose models have been trained and tested on publicly available datasets such as ChestX-ray14 and COVIDx datasets. This may be due to the terms of use for public datasets where most are restricted for research and non-commercial use in the licensing. For example, ChexPert is distributed under a license that allows only non-commercial use [6]. Since many of the publicly available datasets cannot be used for training or validation purposes for commercial tools, the absolute performance of commercial tools cannot be directly compared those developed using open datasets. The datasets used for commercial tools were split based on underlying pathology, age and gender. Despite known bias with AI tools, race subdivision was missing from commercial tools, with no information provided on failure analysis.

Statistical evaluation (AUC, sensitivity, specificity, and accuracy) are used for reporting model performance. Within groups, for example, AI tools for breast, the ground truth varies from product to product, thus comparative analysis between products is difficult. Efforts like the standardized AI use cases [25] developed by the ACR can support AI development by establishing ground truth standards applicable to multiple products.

In terms of types of models being used for the application of AI in radiology, it is clear that both open-source and proprietary tools rely heavily on deep learning methods. Open-source AI tools disclose their models in detail in peer-reviewed publications. On the other hand, few companies were open to disclose the internals of their models. Some companies have published white papers regarding their tools, with a few details about their neural network architectures, but not much more.

While there has been growth of open source tools on deep learning including release of large pre-trained models that are fine tuned for medical imaging, there is no overlap between FDA approved tools and open source tools. Commercial tools do not release their data to the public and do not participate in public challenges. Open source tools at the moment do not have FDA approval. Therefore, scale of implementation of the open source tools is zero, due to lack of FDA approval, while commercial companies are reluctant to share this information. From our sample, companies with large implementation sites were willing to share this information.

We could not find sufficient information on usability and workflow studies of AI tools to the radiologist workflow. Koios DS , working on breast ultrasound AI has performed some usability studies, likely because their system is used by the radiologist during actual image capture when decisions are made. Moreover, with emergence of marketplaces

[6] https://stanfordmlgroup.github.io/competitions/chexpert/

of AI systems available through PACS or reporting vendors, usability and clinical workflow integration is likely going to be separated from the individual AI product, as organizations will purchase a suite of products that fit their current technology stack rather than those that perform best or are most usable.

Even though our survey instrument was based on the CLAIM, STARD2015 and TRIPOD checklists, we recognized that we would not get answers on all the checklist questions during our first conversation with the companies. We understand that companies developing proprietary tools tend to limit discussion of their algorithms and models. Therefore, we developed our modified questionnaire particularly suited for such companies that only included broad-scope question regarding the model. We focused on the evaluation, dissemination, and implementation of the AI tools. We believe this questionnaire can serve as a standardization template for information gathering from companies developing proprietary AI tools for radiology.

Study Limitations

Firstly, our review is restricted by limited information sharing regarding commercial products due to concerns of IP. Nonetheless, we recognize the willingness of companies to share information acquired during the study. We also performed a thorough literature review to obtain additional information on the companies. We narrowed our focus on diagnostic tasks for AI in radiology, and hence omit many applications used in image processing and other steps in the imaging workflow. This is an area for future evaluations, for example to assess the cognitive performance on missing lesions when low dose images are reconstructed with AI methods. The field of AI is rapidly changing, and hence more data will be available. We hope by publishing the modified checklist for evaluating AI products, and also reviewing the open source tools can empower the readers to review new evidence and new products that come into the market.

Conclusion

AI for medical imaging has been characterized by hype, with exaggerated claims of superhuman performance when compared to clinicians [26, 27]. Several recent articles have highlighted the challenge of clinical translation of AI, with most studies focusing on peer review articles and clinical trials reporting [26]. Our review of products adds to the body of evidence, since FDA approval does not mandate peer review, but rather involves retrospective evaluations of AI tools and internal performance review [28, 29]. IP concerns and the efforts required to obtain regulatory approval can affect motivation to improve AI systems in practice, and there is an opportunity for collaborative models including academic-industry partnerships for clinical validation post regulatory approval.

Acknowledgment

We would like to extend our sincere thanks to the exceptional response from the technical and support teams of each of these organizations. Quantib BV (Ory Six), KOIOS Medical Inc. (Uptegrove, Joni; Leo Barinov; Ajit Jairaj), Imaging Biometrics, LLC (Michael Schmainda), iCAD Inc.(Thomas carrington, Senthil Periaswamy), Subtle Medical Inc. (Anna Menhyhart), Vital Images Inc. (Amber St. Martin), RADLogics (Patrick Browning), Zebra Medical Vision (Ayelet Akselrod-Ballin, PH.D), AIDOC (Marlee Long)

Funding

Funding support was received from the U.S. National Science Foundation #1928481 from the Div Of Electrical, Communication & Cyber Systems.

References

[1] *MDR – Article 2 – Definitions – Medical Device Regulation*. https://www.medical-device-regulation.eu/2019/07/10/mdr-article-2-definitions/. (Accessed on 06/06/2020).

[2] Elizabeth West, Simukayi Mutasa, Zelos Zhu, and Richard Ha. “Global Trend in Artificial Intelligence–Based Publications in Radiology From 2000 to 2018”. In: *American Journal of Roentgenology* 213.6 (2019), pp. 1204–1206.

[3] *FDA Cleared AI Algorithms — American College of Radiology*. https://www.acrdsi.org/DSI-Services/FDA-Cleared-AI-Algorithms. (Accessed on 06/06/2020).

[4] Dong Wook Kim, Hye Young Jang, Kyung Won Kim, Youngbin Shin, and Seong Ho Park. "Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers". In: *Korean journal of radiology* 20.3 (2019), pp. 405–410.

[5] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study". In: *PLoS medicine* 15.11 (2018).

[6] Ian Pan, Saurabh Agarwal, and Derek Merck. "Generalizable inter-institutional classification of abnormal chest radiographs using efficient convolutional neural networks". In: *Journal of digital imaging* 32.5 (2019), pp. 888–896.

[7] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. "Dissecting racial bias in an algorithm used to manage the health of populations". In: *Science* 366.6464 (2019), pp. 447–453.

[8] Jérémie F. Cohen et al. "STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration". en. In: *BMJ Open* 6.11 (Nov. 2016). Publisher: British Medical Journal Publishing Group Section: Medical publishing and peer review, e012799. ISSN: 2044-6055, 2044-6055. DOI: 10.1136/bmjopen-2016-012799.

[9] A. Russell Localio and Catharine B. Stack. "TRIPOD: A New Reporting Baseline for Developing and Interpreting Prediction Models". In: *Annals of Internal Medicine* 162.1 (Jan. 2015). Publisher: American College of Physicians, pp. 73–74. ISSN: 0003-4819. DOI: 10.7326/M14-2423.

[10] John Mongan, Linda Moy, and Charles E Kahn Jr. *Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers*. 2020.

[11] Food and Drug Administration. "Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD)-discussion paper." (2019).

[12] Linda Wang and Alexander Wong. "COVID-Net: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images". In: *arXiv:2003.09871 [cs, eess]* (May 2020). arXiv: 2003.09871.

[13] Joseph Paul Cohen, Paul Bertin, and Vincent Frappier. "Chester: A Web Delivered Locally Computed Chest X-Ray Disease Prediction System". In: *arXiv:1901.11210 [cs, q-bio]* (Feb. 2020). arXiv: 1901.11210.

[14] Shih-Cheng Huang et al. "PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging". en. In: *npj Digital Medicine* 3.1 (Apr. 2020). Number: 1 Publisher: Nature Publishing Group, pp. 1–9. ISSN: 2398-6352. DOI: 10.1038/s41746-020-0266-y.

[15] Nan Wu et al. "Deep Neural Networks Improve Radiologists' Performance in Breast Cancer Screening". In: *IEEE Transactions on Medical Imaging* 39.4 (Apr. 2020). Conference Name: IEEE Transactions on Medical Imaging, pp. 1184–1194. ISSN: 1558-254X. DOI: 10.1109/TMI.2019.2945514.

[16] Scott Mayer McKinney et al. "International evaluation of an AI system for breast cancer screening". en. In: *Nature* 577.7788 (Jan. 2020). Number: 7788 Publisher: Nature Publishing Group, pp. 89–94. ISSN: 1476-4687. DOI: 10.1038/s41586-019-1799-6.

[17] Tiancheng He et al. "A Deep Learning–Based Decision Support Tool for Precision Risk Assessment of Breast Cancer". In: *JCO Clinical Cancer Informatics* 3 (May 2019). Publisher: American Society of Clinical Oncology, pp. 1–12. DOI: 10.1200/CCI.18.00121.

[18] Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. "A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction". In: *Radiology* 292.1 (May 2019). Publisher: Radiological Society of North America, pp. 60–66. ISSN: 0033-8419. DOI: 10.1148/radiol.2019182716.

[19] Li Shen, Laurie R. Margolies, Joseph H. Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. "Deep Learning to Improve Breast Cancer Detection on Screening Mammography". en. In: *Scientific Reports* 9.1 (Aug. 2019). Number: 1 Publisher: Nature Publishing Group, p. 12495. ISSN: 2045-2322. DOI: 10.1038/s41598-019-48995-4.

[20] Weicheng Kuo, Christian Hane, Pratik Mukherjee, Jitendra Malik, and Esther L. Yuh. "Expert-level detection of acute intracranial hemorrhage on head computed tomography using deep learning". en. In: *Proceedings of the National Academy of Sciences* 116.45 (Nov. 2019). Publisher: National Academy of Sciences Section: Biological Sciences, pp. 22737–22745. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1908021116.

[21] Mohammad R. Arbabshirani et al. “Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration”. en. In: *npj Digital Medicine* 1.1 (Apr. 2018). Number: 1 Publisher: Nature Publishing Group, pp. 1–7. ISSN: 2398-6352. DOI: 10.1038/s41746-017-0015-z.

[22] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. “ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 1063-6919. July 2017, pp. 3462–3471. DOI: 10.1109/CVPR.2017.369.

[23] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. “PadChest: A large chest x-ray image dataset with multi-label annotated reports”. In: *arXiv:1901.07441 [cs, eess]* (Feb. 2019). arXiv: 1901.07441.

[24] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi, Kanae Kawai Miyake, Mia Gorovoy, and Daniel L. Rubin. “A curated mammography data set for use in computer-aided detection and diagnosis research”. en. In: *Scientific Data* 4.1 (Dec. 2017). Number: 1 Publisher: Nature Publishing Group, p. 170177. ISSN: 2052-4463. DOI: 10.1038/sdata.2017.177.

[25] *TOUCH-AI Directory — American College of Radiology*. https://www.acrdsi.org/DSI-Services/Define-AI. (Accessed on 06/15/2020).

[26] Myura Nagendran et al. “Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies”. In: *bmj* 368 (2020).

[27] Judy W Gichoya, Siddhartha Nuthakki, Pallavi G Maity, and Saptarshi Purkayastha. “Phronesis of AI in radiology: Superhuman meets natural stupidity”. In: *arXiv preprint arXiv:1803.11244* (2018).

[28] *K180647.pdf*. https://www.accessdata.fda.gov/cdrh_docs/pdf18/K180647.pdf. (Accessed on 06/07/2020).

[29] *Artificial Intelligence and Machine Learning in Software as a Medical Device — FDA*. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device. (Accessed on 06/07/2020).