Pediatric Radiology

Artificial Intelligence (AI) Reporting Guidelines: What the Pediatric Radiologist Needs to Know

--Manuscript Draft--

Manuscript Number:	PRAD-D-21-00147R1
Full Title:	Artificial Intelligence (AI) Reporting Guidelines: What the Pediatric Radiologist Needs to Know
Article Type:	Special issue: Artificial intelligence in pediatric radiology
Funding Information:	National Institute for Health Research (NIHR-CDF-2017-10-037) — Dr Owen J Arthurs
Abstract:	There has been an exponential rise in artificial intelligence (AI) research in imaging in recent years. Whilst the dissemination of study data that has the potential to improve clinical practice is welcomed, the level of detail included in early AI research reporting has been highly variable and inconsistent, particularly when compared to more traditional clinical research. However, inclusion checklists are now commonly available and accessible to those writing or reviewing clinical research papers. AI-specific reporting guidelines also exist and include unique requirements, but these can be daunting for radiologists new to the field. Given that pediatric radiology is a specialty faced with workforce shortages and an ever-increasing workload, AI could help by offering solutions to time-consuming tasks thereby improving workflow efficiency and democratizing access to specialist opinion. Pediatric radiologists will therefore be increasingly leading and contributing to AI imaging research, and researchers and clinicians alike should feel confident that the findings reported are presented in a transparent way, with sufficient detail to understand how they apply to wider clinical practice. In this review, we describe two of the most clinically relevant and available reporting guidelines to help increase awareness and engage the pediatric radiologist in conducting AI imaging research. This guide would also be useful for those reading and reviewing AI imaging research and as a checklist with examples of what to expect.
Corresponding Author:	Susan Cheng Shelmerdine, MBBS BSc MRCS FRCR PhD Great Ormond Street Hospital For Children NHS Trust UNITED KINGDOM
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Great Ormond Street Hospital For Children NHS Trust
Corresponding Author's Secondary Institution:	
First Author:	Riwa Meshaka
First Author Secondary Information:	
Order of Authors:	Riwa Meshaka Daniel Pinto Dos Santos Owen J Arthurs Neil J Sebire Susan Cheng Shelmerdine, MBBS BSc MRCS FRCR PhD
Order of Authors Secondary Information:	
Author Comments:	Dear Dr Strouse, Please find attached our invited review article for the upcoming special issue on AI for Pediatric Radiology, commissioned by Prof Amaka Offiah.

	The document provides radiologists a 'blue print' of how to read and report their own AI related research, and we highlight good examples of other authors who have written a clear study protocol and paediatric radiology research paper from the medical literature. We hope that by highlighting some of the notable reporting checklists and giving readers examples of good practice, we can both enhance the quality of peer reviews and submitted publications for AI studies in paediatric radiology. We have included three figures in our manuscript. We have permission from 'Radiology' for use of figure 1 (proof of permission is attached as supplementary material), figure 2 comes from an open access paper (no permission required, authors/paper duly acknowledged) and figure 3 is our own creation. If you have any further queries please do not hesitate to contact ourselves. Best wishes, Susan Shelmerdine (on behalf of all authors)

Title:

Artificial Intelligence (AI) Reporting Guidelines: What the Pediatric Radiologist Needs to Know

Short Title:

Reporting Guidelines for Artificial Intelligence Studies

Authors:

Dr Riwa Meshaka 1, 2, 3

Dr Daniel Pinto Dos Santos 4

Dr Owen J Arthurs, 1, 2, 3

Prof Neil J Sebire 2, 3, 5

Dr Susan C Shelmerdine, 1, 2, 3, 6

Affiliations:

1 Department of Clinical Radiology, Great Ormond Street Hospital for Children, London, UK

2 UCL Great Ormond Street Institute of Child Health, Great Ormond Street Hospital for Children, London, UK.

3 Great Ormond Street Hospital NIHR Biomedical Research Centre

4 Department of Radiology, University Hospital Cologne, Kerpener Str. 62, 50937 Cologne, Germany

5 Department of Pathology, Great Ormond Street Hospital for Children, London, UK

6 Department of Clinical Radiology, St. George's Hospital, London, UK

Corresponding Author/ Main Affiliation:

Dr. Susan Shelmerdine

Department of Clinical Radiology, Great Ormond Street Hospital for Children, London, UK

Email: susan.shelmerdine@gosh.nhs.uk

Telephone: +44 (0) 207 405 9200

Type of manuscript: Invited Review Article

Keywords:

Artificial Intelligence, Children, ~~Radiology,~~ Diagnostic Accuracy ~~Artificial Intelligence~~, Machine Learning, Pediatric radiology, Reporting Guidelines~~, Diagnostic Accuracy~~

ABSTRACT:

There has been an exponential rise in artificial intelligence (AI) research in imaging in recent years. Whilst the dissemination of study data that has the potential to improve clinical practice is welcomed, the level of detail included in early AI research reporting has been highly variable and inconsistent, particularly when compared to more traditional clinical research. However, inclusion checklists are now commonly available and accessible to those writing or reviewing clinical research papers. AI-specific reporting guidelines also exist and include unique requirements, but these can be daunting for radiologists new to the field.

Given that pediatric radiology is a specialty faced with workforce shortages and an ever-increasing workload, AI could help by offering solutions to time-consuming tasks thereby improving workflow efficiency and democratizing access to specialist opinion. Pediatric radiologists will therefore be increasingly leading and contributing to AI imaging research, and researchers and clinicians alike should feel confident that the findings reported are presented in a transparent way, with sufficient detail to understand how they apply to wider clinical practice.

In this review, we describe two of the most clinically relevant and available reporting guidelines to help increase awareness and engage the pediatric radiologist in conducting AI imaging research. This guide would also be useful for those reading and reviewing AI imaging research and as a checklist with examples of what to expect.

Declarations:

Funding information:
OJA is funded by a National Institute for Health Research (NIHR) Career Development Fellowship (NIHR-CDF-2017-10-037). This article presents independent funded research - the views expressed are those of the author(s), and not necessarily those of the NHS, NIHR, MRC, RCR or the Department of Health.

Conflicts of interest:
None of the authors have conflicts to declare.

Availability of data and material:
All relevant information is provided within the manuscript and supplementary material. No new data has been generated by this review article.

Code availability:
Not applicable

Ethics approval:
Ethical approval for this review article was not required.

Consent to participate and publication:
Not applicable

Authors' contributions:
All authors listed in this manuscript fulfil the ICMJE recommendations for authorship. RM and SCS performed the literature review and primary write up of the manuscript. SCS conceived the idea of the study. All authors have had an input in reviewing and editing the final draft of this manuscript.

Acknowledgements:
Not applicable

Artificial Intelligence (AI) Reporting Guidelines: What the Pediatric Radiologist Needs to Know

ABSTRACT

There has been an exponential rise in artificial intelligence (AI) research in imaging in recent years. Whilst the dissemination of study data that has the potential to improve clinical practice is welcomed, the level of detail included in early AI research reporting has been highly variable and inconsistent, particularly when compared to more traditional clinical research. However, inclusion checklists are now commonly available and accessible to those writing or reviewing clinical research papers. AI-specific reporting guidelines also exist and include unique requirements, but these can be daunting for radiologists new to the field.

Given that pediatric radiology is a specialty faced with workforce shortages and an ever-increasing workload, AI could help by offering solutions to time-consuming tasks thereby improving workflow efficiency and democratizing access to specialist opinion. Pediatric radiologists will therefore be increasingly leading and contributing to AI imaging research, and researchers and clinicians alike should feel confident that the findings reported are presented in a transparent way, with sufficient detail to understand how they apply to wider clinical practice.

In this review, we describe two of the most clinically relevant and available reporting guidelines to help increase awareness and engage the pediatric radiologist in conducting AI imaging research. This guide would also be useful for those reading and reviewing AI imaging research and as a checklist with examples of what to expect.

INTRODUCTION

Artificial intelligence (AI) has made an unprecedented impact in radiology, with an ever-increasing number of papers dedicated to new algorithms and uses for machine learning. Approximately 3500 papers related to radiology AI were published in 2020, compared to approximately only 650 in 2015 and less than 250 in 2010. In recent years, there has also been a rise in open access medicine-specific AI journals, for example the Journal of Medical Artificial Intelligence [1]. A number of existing high impact journals have also offered new side publications specifically for AI imaging articles, for example Radiology: Artificial Intelligence [2] and Nature Machine Intelligence [3].

Research in pediatric radiology has shown a relative lag in the rise of AI articles compared to other branches of medicine, possibly due to the overall smaller footprint that pediatric radiology occupies in healthcare (therefore lower return on investment for vendors), greater heterogeneity of cases (both in normality and pathology) and generally fewer case numbers, limiting the acquirement of a large dataset for algorithm training [4]. Despite this, AI algorithms have been validated in a number of studies and could offer a promising solution in managing increasing workload under the strain of workforce shortages. A good example of this is the implementation of BoneXpert™, used routinely in pediatric radiology departments, which automatically assesses bone age in 21 seconds, compared to 165 seconds using the traditional Greulich-Pyle method [5]. Another instance from adult radiology is in breast screening, whereby the application of AI has been shown to potentially reduce radiologist workload by up to 47%, simply by removing normal studies from the workflow [6]. Additionally, AI could be used to aid non-specialized radiologists in diagnosing abnormalities specific to children, for example buckle fractures, or misplaced lines and tubes. For this these reasons, high quality and clinically impactful AI research in pediatric radiology should be encouraged.

For research to be clinically relevant, the methodology and results sections in particular should be communicated in a standard, transparent and reproducible manner. The current wide variability in AI research reporting could lead to misinterpretation, and the use of methods that have not been fully validated or are not generalizable. It also makes the task of reviewing and comparing AI studies challenging and inaccurate. Reporting guidelines that have been adapted for AI research studies have recently been developed such as Checklist for Artificial Intelligence in Medical Imaging (CLAIM) [75] (for diagnostic accuracy assessment), The Standard Protocol Items: Recommendations for Interventional Trials AI extension (SPIRIT-AI) [86-108], and Consolidated Standards of Reporting Trials AI extension (CONSORT-AI) [9-11-13] (for interventional studies). Nevertheless, radiologists who are new to AI research may find it challenging to apply these or be unaware of their existence and importance.

This review therefore aims to guide the pediatric radiologist undertaking AI research, to help report findings in a standardized manner and map good examples of research reporting to the existing guidance. These guidelines should help improve the quality of AI imaging research within the specialty.

AI RESEARCH REPORTING GUIDELINES RELEVENT TO THE PEDIATRIC RADIOLOGIST

Standards for Reporting Diagnostic Accuracy Studies (STARD), 2015 [142] guidance is an accepted method of reporting diagnostic accuracy trials, and an extension to cover AI related studies is still under development [153]. Last year, the Radiological Society of North America (RSNA) devised the CLAIM [75] checklist based on STARD, incorporating AI specifics in the form of six new items, 22 altered pre-existing STARD 2015 items, and 14 original items. Minimum Information for Medical AI Reporting", MINIMAR [164], produced by the American Medical Informatics Association (AMIA) shares many specifics with CLAIM and aims to improve transparency and reduce bias that can easily be introduced to AI research. Table 1 summarizes CLAIM guidance with MINIMAR features mapped to individual items to demonstrate similarity and differences. These are not to be confused with MI-CLAIM [175] (Minimum Information about Clinical Artificial Intelligence Modelling) which are not intended for healthcare professionals. For interventional studies, CONSORT-AI [9-11-13] is adapted from the CONSORT statement with 14 additional items presented in Table 2.

PRINCIPLES FOR CONDUCTING DIAGNOSTIC ACCURACY AI STUDIES IN PEDIATRIC RADIOLOGY: EXAMPLES OF GOOD PRACTICE

As the vast majority of preliminary AI research to date in pediatric radiology has focused on the evaluation of diagnostic accuracy of an AI algorithm, we present worked examples from the literature mapped directly to CLAIM guidance and highlighting MINIMAR guidance where this differs.

Title (CLAIM Item 1)

The title should inform the reader what is being tested and the accuracy measure being assessed, for example sensitivity or accuracy. To inform the reader of the AI nature of the article, the title should mention the type of machine learning algorithm used, for example "convolutional neural network", the most commonly used in medical imaging. This title, "Detection of Traumatic Pediatric Elbow Joint Effusion Using a Deep Convolutional Neural Network" fulfils this criterion by including both the specific AI technology and the term "detection"[186].

Abstract (CLAIM Item 2)

As with non-AI research, the abstract should be a structured summary of the article, including background/purpose, methods, results and conclusions.

Introduction (CLAIM Items 3-4)

The introduction should include the scientific and clinical context, objectives and hypotheses of the study, highlighting the issues that the AI tool will address, including the intended use if validated. A good example of where this was done is from

a recent paper [19~~7~~] explaining how changes in ventricular size in patients with hydrocephalus correlate to clinical status and aid decisions in neurosurgical management. The authors argue that current methods of visual inspection and comparison between different imaging techniques, slice thickness and orientation are not standardized and liable to error and variability between readers. Their AI tool aims to offer a quantitative method to quickly and consistently measure ventricular volume over time which could addresses this clinical need and save radiologist time.

Methods (CLAIM Items 5-32)

The methodology section for an AI research paper differs in many ways to conventional guidance, and most of the new items in the CLAIM guidance relate to this section. ~~Clearly, as with non-AI related research, a study registration number, access to full study protocol and any sources of funding should be stated (Claim items 40 – 42)~~

There are seven key areas to address in the Methods: study design, data, ground truth, data partitions, model, training and evaluation. Table 3 provides three separate examples from the literature and how each of these sections ~~were~~ was addressed.

1. Study design (CLAIM Item 5-6): state whether the data collection was prospective or retrospective, and the goal of the study. For example, to test the feasibility of an AI model, to create a new AI model, or test performance or non-inferiority of a model. For example, Choi et al [20~~18~~] state that the aim was to test feasibility and performance of a dual input machine learning model in detecting pediatric supracondylar fractures.

2. Data (CLAIM Items 7-13): state the sources of data, and whether these were comparable to the data that the AI tool will be exposed to if validated and implemented. State whether ethical approval has been obtained and describe the inclusion and exclusion criteria. If other studies have used the same dataset, explain how this study differs. For example, Larson et al [21~~19~~] obtained 14,036 left hand radiographs taken specifically for bone age assessment from two tertiary children's hospitals. This dataset would be applicable ~~to the~~for future use in assessing bone age as the images were obtained specifically for this purpose. MINIMAR [16~~4~~] advises the inclusion of detailed additional population demographic information which should be presented in the results, such as ethnicity and socioeconomic status, to anticipate potential biases. This allows the reader to assess whether the model could be applied to their own population of patients, or if adjustments would need to be made. Whilst this is the ideal situation, in some studies data protection measures may prevent this additional information from being reported.

   Pre-processing is a vital step in AI research methodology as it standardizes the data into a form that can be read and analyzed by the computer. Pre-processing includes anything that may have altered the image to make it more "readable" to the machine, such as adjusting window settings or image size. For example, Larson et al [21~~19~~]

explain how their images were converted to Portable Graphics Format (PNG), contrast equalized, resolution downsized, and image size cropped in a process of standardization. Explaining the steps taken in pre-processing not only makes the study reproducible, but also reassures the reader that the images used to train the AI have not been manipulated to an unrecognizable or unachievable form. The pre-processed images should still be representative of a clinical dataset. Any inconsistencies that lead to removal of data should be detailed and any de-identification (anonymization or pseudo-anonymization) methods explained. If the data has been split into sub-sets at this stage, this should also be explained.

The specific output variables being measured by the AI should be defined, ideally in the form of radiological Common Data Elements [220] if these exist. In the case of Larson et al [2119], the outcome classifications were normal, advanced, or delayed bone age. In Zhou et al [231], the outcomes were defined as both binary and 3-way: whether a particular tumor was present or absent and which type of tumor of three (astrocytoma, medulloblastoma or ependymoma) was diagnosed. Choi et al [2018] reported the presence or absence of a supracondylar elbow fracture. Other examples could be a particular measurement or presence of a specific sign.

3. Ground Truth (CLAIM Items 14-18): the reference standard (or 'gold' standard) in AI is referred to as "ground truth" and needs to be defined in enough detail to allow replication. Test result categories of the reference standard need to be defined and a measurement of inter- and intra-rater variability given with methods to mitigate these. For example, Larson et al [2119] used reported bone age extracted from radiology reports by an automated script. The original reports were written by expert pediatric radiologists using their standardized technique and therefore the closest approximation available to radiological bone age and a sound representation of ground truth. If there is more than one possible source for ground truth, the study should explain why a particular one was chosen over another (e.g., consensus expert opinion versus pathological result or clinical outcome).

4. Data Partitions (CLAIM Items 19-21): Explain any sample size calculations and how the images were divided into samples. In using an AI tool for detection of a supracondylar elbow fracture, Choi et al [2018] designated 80% of their 1,266 data set to training and 20% to validation. There were two separate test sets. The first contained 258 paired radiographs (frontal and lateral views) from the same institution, separated by time. The second contained 95 paired images from a different institute. The ages of the patients and the ratio of fracture to no fracture in each group were given to demonstrate that the test groups were comparable to the training and validation groups.

5. Model (CLAIM Items 22-24): Explanation of models and their optimization can be complicated, and the aim of the paper should be to provide enough detail to allow reproducibility by other groups. Code can be provided as a supplement if required. The inputs (this should match the pre-processed data) and outputs (this is the outcome variable and should match ground truth) should be described in addition to statements regarding intermediate layers and software libraries, including version numbers, or existing frameworks or models used, as well as modifications made to these. Details should be provided regarding whether random initialization or transfer learning were used to initialize the model. Larson et al [2119] detail the use of a deep residual network with 50 layers and 3.8 x10^9 floating point operations with output including probability score for bone age month and sex. The study used TensorFlow, an open-source machine learning tool implemented by several~~by~~ Google services [24], "Adam" [252] as an optimization algorithm, and pre trained weights converted from an online repository for initialization. The CLAIM guidance suggests avoiding a discussion on specific hardware used, unless a technology assessment is part of the study. MINIMAR [164] includes a statement on how the code and data have been shared and how to access it in the wider research community. Many ~~studies~~ authors choose to include coding data as a supplementary, typically accessed via an online platform (GitHub).

6. Training (CLAIM Items 25-27): This is a section unique to AI related studies and involves describing how the model was trained, including any transformations (or augmentations) applied to the images shown to the AI tool, e.g. if the images were flipped or rotated, how many models, and how a final model was chosen. The included hyperparameters should be described, for example the learning rate schedule, training duration and batch sizes and how a final model was selected. If the final model includes a combination of two models, how these were combined should be documented. In differentiating posterior fossa tumors, Zhou et al [231] replicated their Tree-Based Pipeline Optimization Tool ten times to yield the optimal model which was then chosen for further testing. Both Larson [2119] and Choi [2018] describe in detail the processes of image augmentation used in their studies, including random flips, rotations and zoom. Choi et al [2018] list their hyperparameters including learning rate.

7. Evaluation (CLAIM Items 28-32): This section relates to how the data is analyzed and compared to the reference standard. The researchers should state how performance was measured, which statistical tests and software were used, including measures of uncertainty (95% confidence intervals). Additional areas that are novel to AI which need inclusion are how data are assigned to partitions, measuring robustness, methods for interpretability (e.g., saliency or heat maps) and validation on external data. For example, Zhou et al [231] and Choi et al [2018] both use area under the curve, sensitivity and specificity of their models compared to expert review with confidence

limits stated. Larson et al [21~~19~~] used saliency maps to show the reader which pixels of the image the model was most sensitive to as a heat map (**Figure 1**).

Results (CLAIM Items 33-37)

Researchers will be well versed with presenting participant selection in a flow diagram and baseline demographic data in a table. These should similarly feature in an AI based paper, particularly with regards to the training, validation and testing group splits, tabulated with ground truth. Park et al [26~~3~~] present an example of this in relation to AI for developmental hip dysplasia (DDH) detection (**Table 4~~Figure 2~~**). This is helpful and could be further enhanced with a demographic breakdown and proportions of totals to help the reader compare groups such as in Choi et al [20~~18~~], who present an example alongside a flow chart of the data set sources and partitions (adapted **Table 5 and Figure 2~~3~~**).

Formatted: Font: Bold

Performance metrics for the model should be presented with estimates of diagnostic accuracy and their precision (confidence intervals). Any incorrectly classified cases should be presented and analyzed. Zheng et al [27~~4~~] compared the time it took a radiologist to measure limb length discrepancy on ~~plain film~~conventional radiographs to a machine learning model and found that the radiologist's manual measurement time was a mean of 96 seconds, compared to less than 1 second for the AI. They reported no significant difference in the limb length in the radiology report compared to the AI tool (P>0.05). The authors quote Pearson correlation coefficients and the mean squared error for limb length measurements, and discrepancy between the radiologist and the AI. Errors of more than 2cm are noted in several cases between clinical reports and the AI. These are not discussed individually, rather assigned to "deep learning failure" and a number of possible areas for improvement are ~~discussed~~ presented including diversifying the data, increasing the training numbers and ensuring better image quality. While efforts were made to identify these inaccuracies, incorrectly classified cases should ideally be analyzed in greater detail and discussed individually where possible. This would allow for the identification of targeted areas for improvement.

Discussion (CLAIM Items 38-40)

As with the discussion section of any study, this should include a statement summarizing the results, a comprehensive discussion of ~~its~~ the implications or significance, limitations and any potential bias or uncertainty. The implications for practice (including potential unintended consequences) and the future role for the AI tool should be discussed. Quon et al [28~~5~~] recently published an article describing pediatric posterior fossa tumor detection with a classification tool using T2 weighted MRI. The authors justify their use of T2-weighted images and describe this as a limitation since it may have artificially reduced the performance of the radiology reviewers, who may ordinarily have more than one sequence available before tumor classification ~~was~~ is attempted.

Additional Study Information (CLAIM items 40-42)

As with non-AI related research, a study registration number, access to full study protocol and any sources of funding should be stated (CLAIM items 40 – 42), typically in the Methods section.

Formatted: Font: Bold, Italic

SPECIFICS OF AI INTERVENTION REPORTING

The CONSORT-AI extension to CONSORT [~~9~~-11-13] presented in Table 2 includes many of the details already described in CLAIM [7~~5~~] and MINIMAR [16~~4~~]. There are currently no published examples of such studies in pediatric radiology, however when these emerge it will be important to acknowledge these standards. The main specific items for consideration include an explanation of how the AI was integrated into the clinical workflow, which version of software was used and how the AI output contributed to decision-making.

CONCLUSION

Reporting guidelines specific to AI radiology studies aid researchers in publishing transparent and reproducible work. They allow reviewers and readers to have an expectation of the minimum detail included and ensure a more robust approach to critically appraising papers. Given that AI related studies are still relatively under-represented in pediatric radiology, those conducting AI research in this field may be inexperienced in using these guidelines.

Using the most relevant guidelines to AI diagnostic studies and applying them to examples from the pediatric radiology literature, we have presented a practical guide using "worked examples", ensuring coverage of all the major items, targeting those new to reporting AI related research. Standardized reporting will not only enhance the quality of published work, but ensure reproducibility and comparability for readers and reviewers. As an ever-increasing number of AI studies are published, we must become well versed in basic reporting standards when reading these articles and considering their application to clinical practice.

TABLES

Table 1. Adapted summary table of the components of Checklist for Artificial Intelligence in Medical Imaging (CLAIM) Guidance for reporting Artificial Intelligence (AI) studies on diagnostic accuracy [75] and MINimum Information for Medical AI Reporting (MINIMAR) [164]. Both include similar methodology features. The components specific to MINIMAR are in *italics*.

CLAIM Section	Item	Explanation	Reciprocal MINIMAR Item
Title	1	Identify as AI, specifying category (e.g., deep learning)	
Abstract	2	Structured Summary	
Introduction	3	Background and intended use of AI	
	4	Objectives and hypotheses	
Methods			
Study Design	5	Prospective or retrospective	
	6	Study goal, such as model creation, exploratory, feasibility, noninferiority trial	
Data	7	Data sources	Data source
	8	Eligibility criteria: how, when and where potential participants were identified	Population from which sample was drawn and study setting. Cohort selection (inclusion/exclusion criteria).
	9	Data preprocessing steps	
	10	Selection of data subsets, if applicable	
	11	Definition of data elements	
	12	De-identification methods	
	13	How missing data were handled	Missingness
Ground Truth	14	Definition of ground truth reference standard	Gold Standard
	15	Rationale for choosing the reference standard if alternatives exist	
	16	Source of ground truth annotations and qualifications/preparation of annotators	
	17	Annotation tools	
	18	Measurement of inter- and inter-rater variability and how mitigated, how discrepancies were resolved	
Data Partitions	19	Intended sample size and how it was determined	Data splitting
	20	How data were assigned to partitions (specify proportions)	
	21	Level at which the partitions are disjoint (e.g., image, study, patient, institution)	
Model	22	Model description including inputs, outputs and intermediate layers and connections	Model task – classification or prediction. Model algorithm type.
	23	Software libraries, frameworks and packages	
	24	Initialization of model parameters (e.g., randomization, transfer learning).	
Training	25	Details of training approach including data augmentation, hyperparameters, number of models trained.	List of variables
	26	Methods of selecting the final model	Model or parameter tuning
	27	Ensembling techniques, if applicable	
Evaluation	28	Metrics of model performance	
	29	Statistical measures of significance and uncertainty (e.g., confidence intervals)	
	30	Robustness or sensitivity analysis	
	31	Methods for explainability and interpretability (e.g., saliency maps) and how they were validated	
	32	Validation or testing on external data	
Results			
Data	33	Flow of participants or cases, using a diagram to indicate inclusion and exclusion	
	34	Demographic and clinical characteristics of cases in each partition	*Demographics including age, sex, race, ethnicity, socioeconomic status*
Model Performance	35	Performance metrics for optimal model(s) on all data partitions	Model output
	36	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	
Discussion	38	Study limitations, including potential bias, statistical uncertainty and generalizability	
	39	Implications for practice, including the intended use and/or clinical role	Intended user of the model output
Other information	40	Registration number and name of registry	
	41	Where the full study protocol can be accessed	*Transparency: how code and data are shared with the community.*
	42	Sources of funding and other support; role of funders	

Formatted Table

Table 2. Summary of additional criteria to be included in studies related to Artificial Intelligence (AI) intervention according to the Consolidated Standards of Reporting Trials-Artificial Intelligence (CONSORT-AI) statement. Table adapted from Liu X et al 2020 [9 11-13]. Item explanations within the CONSORT 2010 statement have been omitted.

CONSORT 2010 Section	**CONSORT-AI item**
Title & Abstract	Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model.
	State the intended use of the AI intervention within the trial in the title and/or abstract.
Background & Objectives	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g., healthcare professionals, patients, public).
Participants	State the inclusion and exclusion criteria at the level of participants.
	State the inclusion and exclusion criteria at the level of the input data.
	Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements.
Interventions	State which version of the AI algorithm was used.
	Describe how the input data were acquired and selected for the AI intervention.
	Describe how poor quality or unavailable input data were assessed and handled
	Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users.
	Specify the output of the AI intervention
	Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice.
Harms	Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, justify why not.
Funding	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.

Table 3. Examples from the literature addressing the Checklist for Artificial Intelligence in Medical Imaging (CLAIM) items relating to the methods section of an AI paper.

CLAIM Method Section	Choi et al – pediatric supracondylar fractures [2018]	Larson et al – Bone age [2119]	Zhou et al – Pediatric posterior fossa tumors [231]
Study Design	Retrospective Train and validate an AI, test feasibility and performance	Retrospective Train and validate an AI model, compare performance to reviewers and existing model	Retrospective Test performance of automated machine learning tool, compare it to expert review and feature extraction.
Data	1266 pairs (AP and lateral) from a single institution for training and validation. 258 different pairs from same institution and 95 pairs from different institution for testing. Excluded patients with nonsupracondylar fractures or bone dysplasia Preprocessing: converted to PNG, cropped around a central coordinate, histogram equalized, resized to 200 by 200 pixels. Outcome variable: supracondylar fracture or no fracture.	14036 left hand radiographs taken for bone age assessment from two tertiary children's hospitals Preprocessing: Converted to PNG, contrast equalized, resolution downsized, image cropped Outcome variables: normal, advanced or delayed bone age.	T2, contrast enhanced T1 and ADC images from three tertiary hospitals. 288 patients: 111 medulloblastoma, 107 pilocytic astrocytoma, 70 ependymoma. 3D Slicer used to manually segment tumors, ROIs manually drawn around enhancing and non-enhancing tumor, perilesional edema. Outcome variable: binary (i.e., medulloblastoma or not) and 3-way (medulloblastoma, ependymoma or astrocytoma).
Ground Truth	Two experienced readers reviewed and labelled as supracondylar fracture or non-fracture	Bone age in expert radiology reports extracted automatically. Three additional reviewers used to measure variability.	Histologically confirmed diagnosis.
Data Partitions	80% training + 20% validation, split at random An additional 258 + 95 pairs used for testing which had a similar ratio of fractures: non-fractures.	12611 (90%) training set, 1425 (10%) validation set, split at random. A separate 200 test set from Stanford and 913 test set from a publicly available resource.	70% training + validation and 30% testing set, split at random
Model	Keras (Python-based) run on top of ~~Google~~ TensorFlow. Each image input to 2 identical ResNet-50 models. Output: prediction value between 0 and 1 for the two labels (fracture or no fracture).	50-layer deep residual network Output = probability score for bone age month and sex ~~Google~~ TensorFlow Adam for optimization Pretrained weights for initialization	Radiomics feature selection by a machine learning expert verses Tree-Based Pipeline Optimization Tool (TPOT) – automatically optimized pipeline based on input. Outputs: binary (i.e., medulloblastoma or not) and 3-way.
Training	Images augmented using a data generator with random combinations of flips, rotations and zoom. Hyperparameters: categorical cross-entropy as the loss function, specified learning rate, decay and momentum, 31625 iterations in 4 batches.	Flipped, contrast-adjusted and cropped images presented Model tested on training set range of sizes 1558-12611	TPOT pipeline replicated 10 times to yield 10 models. Best chosen and testing set applied.
Evaluation	Model compared to 95 test pairs read by radiologists. Area under the curve, sensitivity, positive and negative predictive values measured and compared.	Root mean square and mean absolute difference between the model estimates and reference standard using 200 test images. Compared using paired T and F test respectively. Chi squared test was used to assess the differences between the reviewers and the model. Saliency maps presented.	Area under the curve, accuracy, sensitivity and specificity of the automatic machine learning model verses optimized feature selection method verses expert MR review by two neuroradiologists. 95% confidence intervals and P values calculated.

Formatted: Line spacing: single

Table 4. An example of how to present proportions or normal versus abnormal cases after data splitting in a multi-centre dataset. Park et al [26] present numbers of developmental hip dysplasia (DDH) cases in each split set (training, validation and test sets) to demonstrate that each is proportionally representative of the entire dataset used. PNUYH = Busan National University Yansan Hospital, SNUBH = Seoul National University Bundang Hospital, SNUH = Seoul National University Hospital.

Reproduced (open access) from [26]

Hospitals	Total	Training Set		Validation Set		Test Set	
		Normal	DDH	Normal	DDH	Normal	DDH
SNUH	3433	2406	341	300	43	300	43
SNUBH	1036	800	32	97	5	97	5
PNUYH	607	452	19	65	3	66	2
Total	5076	3658	392	462	51	463	50

Formatted: Font: Italic

Formatted: No underline

Formatted: Centered

Formatted Table

Formatted: Font: Not Bold, No underline

Formatted: Centered

Formatted: Font: Not Bold, No underline

Formatted: Centered

Formatted: Font: Not Bold, No underline

Formatted: Centered

Formatted: Font: Not Bold, No underline

Formatted: Centered

Table 5. Example of how to present a table (a) of demographic breakdown between training, validation and test groups using fictional data. Layout has been adapted from the table by Choi et al [20].

Demographic	**Training Set (n=1000)**	**Validation Set (n=200)**	**Test Set (n=200)**
Age			
0-4	500 (50%)	100 (50%)	100 (50%)
5-9	250 (25%)	50 (25%)	50 (25%)
>9	250 (25%)	50 (25%)	50 (25%)
Label			
Disease	200 (20%)	40 (20%)	40 (20%)
No Disease	800 (80%)	160 (80%)	160 (80%)

Formatted: Font: Bold
Formatted: Centered
Formatted: Font: Bold
Formatted: Centered
Formatted: Centered
Formatted: Centered
Formatted: Font: Bold
Formatted: Centered
Formatted: Centered
Formatted: Centered

Figure Legends

Figure 1

Three examples of saliency maps of paediatric bone hand radiographs from Larson et al [21~~19~~], where an AI tool was trained to evaluate the bone age in children. The saliency map demonstrates the regions of the radiograph (in red and yellow colour) where the AI tool appeared to gather the most useful radiographic information in generating it~~'~~s output variable (i.e. bone age), superimposed on the original radiographs in three male patients aged 4 years (a), 15 years (b) and 17 years (c). ~~.~~

Reproduced with permission [21]~~from Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP (2018) Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology. Apr;287(1):313-322. doi: 10.1148/radiol.2017170236. Epub 2017 Nov 2. PMID: 29095675.~~

~~**Figure 2**~~

~~An example of how to present proportions or normal versus abnormal cases after data splitting in a multi-centre dataset. Park et al [23] present numbers of developmental hip dysplasia (DDH) cases in each split set (training, validation and test sets) to demonstrate that each is proportionally representative of the entire dataset used. PNUYH = Busan National University Yansan Hospital, SNUBH = Seoul National University Bundang Hospital, SNUH = Seoul National University Hospital.~~

~~*Reproduced from Park HS, Jeon K, Cho YJ, Kim SW, Lee SB et al (2020) Diagnostic Performance of a New Convolutional Neural Network Algorithm for Detecting Developmental Dysplasia of the Hip on Anteroposterior Radiographs. Korean J Radiol. Nov 26 (open access).*~~

Figure 2

An example flow-chart of data sources and partitioning where fictional data for algorithm development and training is acquired from one centre, and then validated both internally on a temporal, prospective dataset as well as an external dataset. Layout has been adapted from the flowchart by Choi et al [20].

~~**Figure 3**~~

~~Example of how to present a table (a) of demographic breakdown between training, validation and test groups using fictional data and (b) an example flow-chart of data sources and partitioning where data for algorithm development and~~

~~training is acquired from one centre, and then validated both internally on a temporal, prospective dataset as well as an external dataset. Layout has been adapted from the table and flowchart by Choi et al [18].~~

References

1 Editorial Office (2018) Foreword. Journal of Medical Artificial Intelligence 1:1.

2 Charles E. Kahn J (2019) Artificial Intelligence, Real Radiology. Radiology: Artificial Intelligence 1:e184001

3 Editorial (2019) More than machines. Nature Machine Intelligence 1:1-1

4 Davendralingam N, Sebire NJ, Arthurs OJ, Shelmerdine SC (2020) Artificial intelligence in paediatric radiology: Future opportunities. Br J Radiol. 10.1259/bjr.20200975:20200975

5 Booz C, Yel I, Wichmann JL et al (2020) Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method. Eur Radiol Exp. 28;4(1):6

6 Rodriguez-Ruiz A, Lång K, Gubern-Merida A et al (2019) Can we reduce the workload of mammographic screening by automatic identification of normal exams with artificial intelligence? A feasibility study. European radiology, *29*(9), 4825–4832

75 Mongan J, Moy L, Charles E. Kahn J (2020) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiology: Artificial Intelligence 2:e200029

86 Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26:1351-1363

97 Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2:e549-e560

108 Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. Bmj 370:m3210

119 Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26:1364-1374

120 Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet Digit Health 2:e537-e548

131 Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. Bmj 370:m3164

142 Bossuyt PM, Reitsma JB, Bruns DE et al (2015) STARD 2015: An Updated List of Essential Items for Reporting Diagnostic Accuracy Studies. Radiology 277:826-832

153 Sounderajah V, Ashrafian H, Aggarwal R et al (2020) Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med 26:807-808

164 Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH (2020) MINIMAR (MINimum Information for Medical AI Reporting): Developing reporting standards for artificial intelligence in health care. Journal of the American Medical Informatics Association 27:2011-2015

175 Norgeot B, Quer G, Beaulieu-Jones BK et al (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. Nat Med 26:1320-1324

186 England JR, Gross JS, White EA, Patel DB, England JT, Cheng PM (2018) Detection of Traumatic Pediatric Elbow Joint Effusion Using a Deep Convolutional Neural Network. AJR Am J Roentgenol 211:1361-1368

Formatted: Font: 11 pt

Formatted: Font: (Default) Arial, 10 pt

Formatted: Font: (Default) Arial, 10 pt

Formatted: Normal

Formatted: Font: Not Italic

Formatted: Font: (Default) Times New Roman, 12 pt

1~~9~~7 Quon JL, Han M, Kim LH et al (2020) Artificial intelligence for automatic cerebral ventricle segmentation and volume calculation: a clinical tool for the evaluation of pediatric hydrocephalus. J Neurosurg Pediatr. 10.3171/2020.6.Peds20251:1-8

20~~1~~8 Choi JW, Cho YJ, Lee S et al (2020) Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. Invest Radiol 55:101-110

21~~1~~9 Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP (2018) Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology 287:313-322

2~~0~~ Radiological Society of North America, RSNA; American College of Radiology, ACR (2021) RadElement: Common Data Elements. Available via https://RadElement.org. Accessed 10 February 2021

2~~1~~ Zhou H, Hu R, Tang O et al (2020) Automatic Machine Learning to Differentiate Pediatric Posterior Fossa Tumors on Routine MR Imaging. AJNR Am J Neuroradiol 41:1279-1285

24 Abadi M, Agarwal A, Barham P et al (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

2~~2~~ Kingma D, Ba J (2015) Adam: A Method for Stochastic Optimization. arXiv. arXiv:1412.6980

2~~3~~ Park HS, Jeon K, Cho YJ et al (2020) Diagnostic Performance of a New Convolutional Neural Network Algorithm for Detecting Developmental Dysplasia of the Hip on Anteroposterior Radiographs. Korean J Radiol. 10.3348/kjr.2020.0051

2~~4~~ Zheng Q, Shellikeri S, Huang H, Hwang M, Sze RW (2020) Deep Learning Measurement of Leg Length Discrepancy in Children Based on Radiographs. Radiology 296:152-158

2~~5~~ Quon JL, Bala W, Chen LC et al (2020) Deep Learning for Pediatric Posterior Fossa Tumor Detection and Classification: A Multi-Institutional Study. AJNR Am J Neuroradiol 41:1718-1725

Formatted: Font: (Default) Arial, 10 pt

Formatted: HTML Preformatted, Line spacing: At least 15 pt

Formatted: Font: (Default) Arial, 10 pt

Formatted: Font: (Default) Arial, 10 pt

Formatted: Font: (Default) Arial, 10 pt

Formatted: Font color: Custom Color(RGB(55,71,79))

Figure 1a

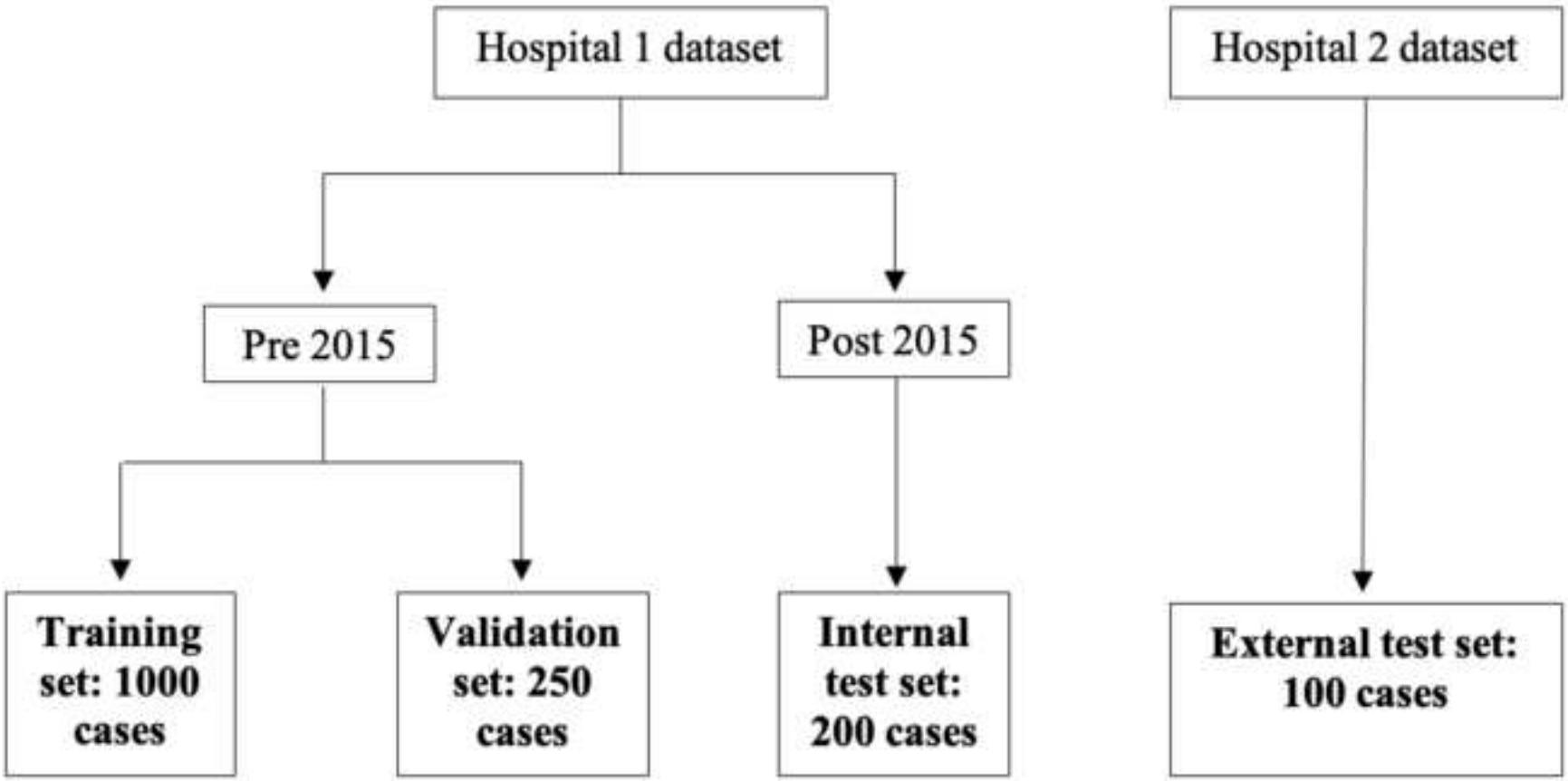
Click here to access/download;Figure;Figure 1a - permission granted.png

Figure 1b

Figure 1c

Click here to access/download;Figure;Figure 1c - permission granted.png

Figure 2



Dear Dr Strouse,

Thank you for your review and comments on our submitted manuscript, "Artificial Intelligence (AI) Reporting Guidelines: What the Pediatric Radiologist Needs to Know"

Please find attached a revised title page and manuscript (tracked and clean copies) with a point-by-point response below.

Kind regards,

Dr Susan Shelmerdine

Editor's comments:

This is a well written and very informative introduction to reporting guidelines for artificial intelligence (AI) manuscripts and a terrific and much needed contribution to the special issue on AI. This is excellent guidance for those doing work in the AI arena and hoping to publish results.

The reviewers have a few suggestions. Please address these carefully in your revision. In addition:

1. Keywords, please add Children, Pediatric radiology; please alphabetize

This has been amended on the Title Page.

2. Line 107 – please put CLAIM in all caps, as elsewhere.

This has been amended and this line has been moved to after the discussion section, as per Reviewer #1 comment.

3. Tables – please spell out abbreviations at first use in the legends; as noted, there is inconsistent use of periods (hard stops) – suggest deleting all

This has been amended in the tables and table legends.

4. Figure 1 – although this is taken from another source, please divide into three separate figure parts (a, b, c) and image files. In the legend, please distinguish between what is seen on each image. If possible, please include the age and gender of each subject illustrated.

Figure 1 has been separated into three and the legend has been amended.

5. Figure 2 – rather than an image of the table from another source, please recreate as a table in editable form in Word. This would be Table 4 rather than Figure 2.

Figure 2 has been recreated as Table 4

6. Figure 3a – please submit in editable form in Word. This would be Table 5 rather than Figure 3a. Figure 3b becomes Figure 2 – please only capitalize the first word in each box.

Figure 3a has been recreated as Table 5 and Figure 3b is now Figure 2 with only first word capitalized in each box.

7. Figure 1 and Figure 2 legends – it is not necessary to give the full citation; "Reproduced with permission [19]" suffices.

This has been amended.

Reviewers' comments:

Reviewer #1

An excellent paper, thank you. I have just a few comments

1. Line 103: The heading states, "Methods (CLAIM Items 5-32)" and yet in Line 107, under that heading, CLAIM items 40-42 are mentioned. It would be preferable to address items 5 to 32 and all CLAIM items in chronological order. Typically study registration number, protocol and sources of funding appear at the end of the Methods section, so this sentence could be moved to allow CLAIM items to be discussed in chronological order. (Note missing full stop after "40-42)"

This has been amended and moved to after the discussion section to follow chronological order.

2. All examples are of good practice, I think the paper would be improved by contrasting with a few examples of poor practice

Thank you – this idea was discussed at the onset of writing this paper and after deliberation, it was decided that citing good examples would be the best way to explain best practice, rather than drawing attention to poor practice, which could have negative consequences to the authors in this new field (in which we are all learning).

3. Lines 234-235: Please be more explicit, is the failure of Zheng et al to discuss individually a good thing or not?

It would have been preferable for the authors to discuss their incorrectly classified cases individually - a line has been added to clarify this.

4. Line 185: Is this correct? Not sure what is meant by, "…learning tool by Google, Adam [22] as an optimization…" Furthermore, Adam is not the first author of Reference 22

This has been clarified to explain that the TensorFlow model is used by Google and a reference has been added (22). "Adam" is the name of the optimization tool rather than the name of the author. Quotation marks have been added to emphasise this.

5. Please check that the citation for Reference 1 is complete

This has been amended.

6. Typos

a. Line 110: change "were" to "was"
b. Line 123: change "to the" to "for"
c. Line 189: change "studies" to "researchers" or "authors"
d. Line 189: change "chose" to "choose"
e. Line 229: change "plain film" to "conventional radiographs"
f. Line 241: change "its" to "the"
g. Line 247: change "was" to "is"
h. Line 348: change "it's" to "its"
i. Table 1: Some fields have a full stop at the end of a sentence, most do not. Please be consistent
j. Table 2: As for Table 1

k. Table 3: As for Tables 1 and 2

Many thanks, these have been amended.

Reviewer #2

1. Line 34: Is the lag in pediatric-specific AI articles also related to overall smaller footprint that pediatric radiology occupies in health care and lower return on investment for industry/vendors?

Many thanks, this excellent point has been added.

2. Line 38: Do any current studies directly study the correlation between an AI tool and potential reduction in workload for the radiologist? Or, will these tools allow nonspecialized radiologists to handle the pediatric radiology exams with more confidence?

Examples of workload reduction and a line about allowing nonspecialized radiologists to interpret pediatric radiology exams have been added.