HHS Public Access
Author manuscript
Emerg Radiol. Author manuscript; available in PMC 2023 November 12.

Published in final edited form as:
Emerg Radiol. 2023 June ; 30(3): 251–265. doi:10.1007/s10140-023-02120-1.

Artificial intelligence CAD tools in trauma imaging: a scoping review from the American Society of Emergency Radiology (ASER) AI/ML Expert Panel

David Dreizin[1], **Pedro V. Staziaki**[2], **Garvit D. Khatri**[3], **Nicholas M. Beckmann**[4], **Zhaoyong Feng**[5], **Yuanyuan Liang**[5], **Zachary S. Delproposto**[6], **Maximiliano Klug**[7], **J. Stephen Spann**[8], **Nathan Sarkar**[9], **Yunting Fu**[10]

[1]Department of Diagnostic Radiology and Nuclear Medicine, R Adams Cowley Shock Trauma Center, University of Maryland School of Medicine, Baltimore, MD, USA

[2]Cardiothoracic Imaging, Department of Radiology, Larner College of Medicine, University of Vermont, Burlington, VT, USA

[3]Department of Radiology, University of Washington School of Medicine, Seattle, WA, USA

[4]Memorial Hermann Orthopedic & Spine Hospital, McGovern Medical School at UTHealth, Houston, TX, USA

[5]Epidemiology & Public Health, University of Maryland School of Medicine, Baltimore, MD, USA

[6]Division of Emergency Radiology, Department of Radiology, University of Michigan, Ann Arbor, MI, USA

[7]Sheba Medical Center, Ramat Gan, Israel

[8]Department of Radiology, University of Alabama at Birmingham Heersink School of Medicine, Birmingham, AL, USA

[9]University of Maryland School of Medicine, Baltimore, MD, USA

[10]Health Sciences and Human Services Library, University of Maryland, Baltimore, Baltimore, MD, USA

Abstract

Background—AI/ML CAD tools can potentially improve outcomes in the high-stakes, high-volume model of trauma radiology. No prior scoping review has been undertaken to comprehensively assess tools in this subspecialty.

Purpose—To map the evolution and current state of trauma radiology CAD tools along key dimensions of technology readiness.

David Dreizin daviddreizin@gmail.com.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10140-023-02120-1.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Methods—Following a search of databases, abstract screening, and full-text document review, CAD tool maturity was charted using elements of data curation, performance validation, outcomes research, explainability, user acceptance, and funding patterns. Descriptive statistics were used to illustrate key trends.

Results—A total of 4052 records were screened, and 233 full-text articles were selected for content analysis. Twenty-one papers described FDA-approved commercial tools, and 212 reported algorithm prototypes. Works ranged from foundational research to multi-reader multi-case trials with heterogeneous external data. Scalable convolutional neural network–based implementations increased steeply after 2016 and were used in all commercial products; however, options for explainability were narrow. Of FDA-approved tools, 9/10 performed detection tasks. Dataset sizes ranged from < 100 to > 500,000 patients, and commercialization coincided with public dataset availability. Cross-sectional torso datasets were uniformly small. Data curation methods with ground truth labeling by independent readers were uncommon. No papers assessed user acceptance, and no method included human–computer interaction. The USA and China had the highest research output and frequency of research funding.

Conclusions—Trauma imaging CAD tools are likely to improve patient care but are currently in an early stage of maturity, with few FDA-approved products for a limited number of uses. The scarcity of high-quality annotated data remains a major barrier.

Keywords

Radiology; Imaging; Emergency; Trauma; Emergency radiology; Artificial intelligence; Machine learning; Computer-aided detection; Scoping review

Introduction

Artificial intelligence and machine learning (AI/ML)–based automated computer-aided detection and diagnosis (CAD) tools that detect, classify, grade, quantify, risk-stratify, or prognosticate injury can potentially improve patient outcomes in the high-stakes, safety–critical, high-volume model of trauma care. Dedicated round-the-clock ER/trauma staff are rarely available outside of busy level I referral centers, and such tools may stand to have the greatest impact on patient outcomes in locations with fewer resources and human expertise [1]. In addition, CAD tools could improve turnaround times, decrease the rate of significant errors, and reduce the often substantial inter- and intra-observer variabilities of classification systems such as AO fracture and American Association for the Surgery of Trauma (AAST) organ grading systems [2–13].

CAD tools: concepts and trends

CAD development emerged in the 1980s using artificial intelligence solutions relying on painstaking pre-defined task-specific hand-crafted feature engineering and often did not generalize or scale well to new data, with few examples of commercialization and widespread adoption beyond mammography [14]. The year 2012 marked a watershed year for supervised feature representation machine learning (i.e., deep learning). Advances in inexpensive graphics processing units (GPUs) with rapid parallel computing architectures and the superior classification performance of the AlexNet convolutional neural network

(CNN) on the large-scale ImageNet dataset in 2012 [15] were followed by the development and open-sourcing of progressively deeper networks [16], encoder-decoder networks (U-Net) for medical imaging segmentation tasks [17], and the introduction of "transfer learning" (pre-training on non-medical images and fine-tuning to medical tasks) in 2015–2016 [18]. The resulting "second boom" in radiology CAD tool development [19] corresponded with an exponential global growth trend in AI/ML publications [19, 20]. Radiology is a data-driven field that was quick to adopt digitized RIS/PACS systems [20, 21] and has been especially well-positioned to catch the tailwinds of new information technologies, with disproportionate growth in FDA-approved AI/ML tools in recent years [22]. Approximately 58% of FDA-approved AI/ML tools are in the radiology domain [22].

The FDA Center for Devices and Radiological Health currently classifies tools that stage, diagnose, or triage pathology as CADt (triage), CADe (detection), CADx (diagnosis), or CADe/x (both detection and diagnosis). IPQ (image processing and quantification) is another common designation for software that is not disease-specific [22]. As many as 99% of AI/ML CAD tools are regulated through the 510 k or de novo software as medical device (SaMD) pathway, and evaluated based on equivalence to existing devices, or whether devices work as intended without major risks [23]. While experts are optimistic regarding mainstream adoption of deep learning algorithms into future clinical practice [24, 25], implementation into practice is currently not widespread [26].

Traumatic injury represents a leading cause of death and disability in patients under the age of 45 with disproportionate effects on quality of life, years of lost productivity, and economic consequences [27, 28]. Although trauma imaging has some overlap with the body region-specific subspecialties, it remains a subfield within the understudied and underfunded field of emergency radiology, and the evolution of AI/ML in this domain has not been well explored.

The American Society of Emergency Radiology convened the ASER AI/ML Expert Panel in 2020 to develop a better understanding of research and development (R&D) in our field and align future clinical and research priorities with the needs of our community. There are still numerous bottlenecks to adopting these tools, and it is important to ascertain the degree of growth in the subspecialty and the extent to which such growth has led to clinical progress [29]. The R&D pipeline of AI/ML tools spans the process of dataset selection, demonstration of proof of concept, productization and deployment, validation of diagnostic performance, and outcomes research [30].

The objective of this systematic scoping review was to synthesize the existing literature and map the evolution and current state of maturity of trauma radiology AI/ML CAD tools along key dimensions of AI/ML technology readiness.

Methods

The scoping review was conducted using a systematic approach according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards and guidelines and the Arksey and O'Malley framework [31]. A scoping review is a structured evidence synthesis method appropriate for mapping the extent, range, and characteristics of research in an area with a broad, complex, and heterogenous nature [32]. The Arksey and O'Malley methodology involves five steps: (1) formulating the research question, (2) identifying relevant studies, (3) study selection, (4) charting the data, and (5) collating, summarizing, and reporting results.

Formulating the research question

To our knowledge, no prior scoping or other structured review has been undertaken to map the level of maturity of automated CAD tools in trauma imaging. The American Society of Emergency Radiology AI/ML Expert Panel created a working group to map the current state and evolution of this field—from foundational or clinical-translational research through post-market validation and outcome studies.

The research question was initially formulated through working group meetings and a cursory exploration of the published literature. Dimensions of AI/ML maturity, reflecting the current state of evolution, adoption, and readiness were explored. Key elements of each theme are described in detail below.

Key dimensions of AI/ML tool maturity

In 2018, the NIH, RSNA, ACR, and the Academy for Radiology and Biomedical Imaging Research held a multidisciplinary workshop to identify major priorities for accelerating foundational and clinical translational AI/ML radiology research [33, 34]. Key themes emphasized in this workshop and a variety of subsequent influential position papers, commentaries, and editorials include the obstacle of data scarcity and the need for high-throughput and high-quality data curation; development of novel architectures to address complex imaging problems; ensuring trustworthiness and fairness of AI/ML methods through interpretability and explainability; performance validation study design that emphasizes robustness and generalizability; a greater focus on user experience and acceptance; and research that emphasizes the value of AI/ML tools with respect to patient care and outcomes. We also consider funding priorities as an important driver of AI/ML research. A detailed discussion of dimensions of AI/ML maturity is provided in Appendix 1.

Identifying relevant studies

In collaboration with a medical research librarian, we conducted a systematic literature search of PubMed, Embase, and Scopus for abstracts using the following: (1) different combinations of keywords pertaining to traumatic injury, radiology, and AI/ML and (2) commercial tool and vendor names extracted from the FDA directory of AI/ML Enabled Medical Devices and list of RSNA 2021 annual meeting exhibitors. Vendor screening was conducted by two independent researchers and arbitrated by a third after contacting vendors for clarification as needed. Keywords tailored to the specific constraints and requirements of each database are listed in Appendix 2. We used a 10-year collection period informed by the expert panel's baseline familiarity with literature using hand-crafted techniques for trauma-related tasks and the first successful implementation of convolutional neural networks on ImageNet in 2012 [15]. Our last search was performed on June 6, 2022.

Study selection

Two levels of review were conducted- a title and abstract screening and a full-text eligibility review. All abstract and full-text reviewers were radiologist members of the ASER AI/ML Expert Panel. Abstract records were screened by two reviewers with disagreements arbitrated by a third. Four members contributed to full-text eligibility screening. All full-text exclusions were determined after review by at least two of the four investigators with arbitration by another.

Our scope was limited to automated computer vision–based CAD tools or prototypes for traumatic injury detection, diagnosis, characterization, triage, risk stratification, or outcome prediction. Works that did not specifically address automated image interpretation that could be expected to fall into one of the FDA CAD categories were not eligible. Excluded papers described the following: image reconstruction or enhancement tasks such as super-resolution imaging; semi-automated or manual quantitative techniques; clinical predictors such as trauma mechanism, baseline demographics, ICD-9 codes, or imaging report keywords with no medical image analysis; segmentation of anatomic structures using a trauma cohort without traumatic injury detection or diagnosis-related tasks (as detailed above); and methods relevant exclusively to the elective and not emergency setting. Abstract-only records and foreign language texts were also excluded, the latter due to the prohibitive cost and time of translation.

Charting the data

After organizing papers by modality and body region/system, full-text papers were charted according to key issues, themes, and characteristics of interest in Microsoft Excel. Key themes included dataset size and curation, algorithm novelty, methods of explainability, validation study methodology, outcomes research, and funding sources. The scoping review method is described by Arksey and O’Malley as an iterative process requiring researchers to adapt their approach to ensure that literature is covered in a comprehensive way [31]. Developing a detailed understanding of the key dimensions of AI/ML maturity required an unstructured but thorough review exploring and re-exploring key position papers during the formulation and gathering of data elements. Thus, if a salient topic became apparent during data gathering, relevant guiding documents were reviewed (see Appendix 1 for detailed discussion), with subsequent fallback and re-review of all included papers for the given element. A description of included data elements is provided in Tables 1 and 2.

Collating, summarizing, and reporting results

We performed simple descriptive statistics using the data elements described above and produced charts and tables to illustrate and map key observations and trends.

Results

We identified 4386 records through a search by AI/ML-related keywords and 231 records through search by vendor and proprietary tool keywords, resulting in a total of 4617 records. After 565 records were excluded as duplicates, 4052 unique records were screened for inclusion. Of those, 3756 records were excluded during screening, and 66 records were excluded during full-text eligibility review. Three records were added following a review of citations. In total, 233 full-text articles were included for content analysis and charting. Our PRISMA flowchart is depicted in Fig. 1.

Characteristics of included studies are summarized in Table 1. Modalities employed included radiography ($n = 104$), computed tomography (CT) ($n = 94$), magnetic resonance imaging (MRI) ($n = 29$), ultrasound (US) ($n = 4$), and DEXA scan ($n = 2$).

Data curation

Nineteen of studies (44/233) reported dataset sizes of less than 100 patients, 42% (99/233) reported using between 100 and 1000 patients, 19% (45/233) in 1000–5000 patients, 14% (32/233) in 5000–50,000 patients, 1% (3/233) in 50,000–500,000 patients, and < 1% (2/233) in more than 500,000 patients. Dataset size was related to the imaging modality and anatomical region/system, with the largest datasets of more than 500,000 patients for chest radiography [35, 36], followed by musculoskeletal radiography with 314,866 patients [37], and a head CT dataset with 313,318 exams [38]. Abdominopelvic trauma imaging papers included the smallest range of dataset sizes with fewer than 100 to 253 patients [39] followed by chest/cardiovascular cross-sectional imaging with fewer than 100 to 778 patients [40]. Most studies used siloed (i.e., non-public) single center data; however, there was a trend toward increased use of both siloed multicenter data and public data over time (Fig. 2).

Types of CAD tools

In total, 148 of the 233 included papers (64%) reported CAD tools or prototypes for detection tasks, 14 (6%) for classification tasks, 27 (12%) for combined detection and classification, 31 (13%) with segmentation/3D quantitative visualization, 8 (3%) employing automated caliper measurements, and 5 papers described computer vision AI/ML for risk stratification and prognostication. As of last search, 212 papers reported CAD algorithm prototypes, and 21 of the 233 included papers (9%) evaluated 10 commercial CAD tools with FDA approval (Fig. 3). In total, 156 papers (67%) report on musculoskeletal (MSK)-related tools, 49 papers on neuroimaging tools (21%), 15 papers (6%) chest and cardiovascular tools, and 13 papers (6%) abdominopelvic imaging tools (Suppl Table 1). The most common MSK use cases reported included detection of extremity, spine, rib, hip, and pelvic fractures (111/156, 71%), combined fracture detection and classification (26/156, 17%), and fracture classification alone (12/156, 8%). Only six papers (4%) described quantitative tools using segmentation masks or electronic caliper measurements.

Detection tools accounted for 49% of neuroimaging papers (24/49), with use cases including ICH detection on CT in 15 of these studies (63%). The remainder were concerned with detection of midline shift, cerebral micro-bleeds, white matter injury, spinal cord injury, and calvarial or facial fracture. Of the tools, 19/49 (39%) involved quantitative imaging, and 4 tools (8%) provided prognostic information only.

Seven cardiovascular/chest papers (47%) described detection tools for pneumothorax, effusions, and opacities, and one described detection of aortic dissection on CT. Seven papers described quantitative visualization tools (47%) for pathology including contusions and pneumothorax on CT and pneumothorax on radiographs. Eight of 13 abdominopelvic papers (62%) reported CT quantitative imaging tools for bleeding and solid organ injury, and the remaining five described detection tools for solid organ injury and active bleeding on CT, pneumoperitoneum on plain radiographs, and free fluid on FAST.

Commercial FDA-approved tools

All 21 studies using commercial tools with FDA approval at last search were published between 2018 and 2022. FDA-approved products included Aidoc ICH [41–47], QureAI qER [38], and RAPID [48] for CT intracranial hemorrhage detection, triage, and notification; Annalise Enterprise CXR [36, 49], which detects rib, scapular, and humeral fractures, pneumothorax, and pneumomediastinum among numerous other findings unrelated to trauma on chest radiographs; Aidoc C-Spine CSF [50, 51] and Aidoc RibFx [52] for CT cervical spine and rib fracture detection, triage, and notification; Gleamer BoneView [53–55], AZMed Rayvolve [56], and Imagen FractureDetect FX [37] for plain radiograph detection of appendicular and some axial fractures (e.g., pelvis); and Siemens AI Rad Companion, which includes a module for vertebral fracture detection and classification [57]. All tools employed CNN architectures. Reported methods of explainability included activation maps for Aidoc ICH and CSF tools; arrow annotations for Aidoc RibFx; bounding boxes for Gleamer BoneView, AZMed Rayvolve, and Imagen FractureDetect FX; segmentation contours or masks for RAPID and Annalise.ai; and segmentations with vertebral height mid-sagittal electronic caliper measurements for Siemens AI Rad Companion. Qure.ai was noted to provide slice-level confidence scores for ICH detections.

Only the Siemens tool, a quantitative visualization method that outputs midsagittal vertebral height measurements, can be considered to provide intrinsic interpretability for a multistage task, as caliper measurements are used in practice to determine the Genant vertebral fracture grade [58]. The other methods, all of which perform a single-stage task (namely, detection), use forms of post hoc explainability.

Although user interfaces may provide the option of reporting false positives or negatives, none of the tools was reported to have interactive capability. An example of such functionality might include the ability to adjust caliper measurements to modulate Genant grade. Furthermore, no paper assessed subjective dimensions of user acceptance such as satisfaction with the user interface, trust, or effects related to workload, effort, or frustration. No paper described novel algorithm development, nor advanced methods for data labeling quality, throughput, or augmentation; however, this is wholly expected for productized proprietary tools far along the technology readiness pipeline. Ground truth was determined by independent readers with arbitration of discrepancies in six papers [37, 38, 45, 54, 56, 57] and using an IoU cutoff > 25% for box proposals in 2 studies [53, 55]. In another [36], 7 radiologists labeled the test dataset, and Dawid-Skene voter aggregation modeling was used to determine ground truth. A single study provided metrics for level of agreement between independent experts [45]. Consensus reads were used in 3 studies [38, 48, 50]. In 5 studies [41, 43, 44, 46, 51], two of which indicate failure mode analysis as the primary purpose [46, 51], only cases flagged by the tool were reviewed further by expert readers without a uniform ground truth process to confirm validity of true positive and negative assignments (i.e., concordant results were assumed to be correctly classified).

Radiology reports alone were used for ground truth in 3 papers [49, 52, 59], and limited information on ground truth methodology was provided in one [42]. The largest validation set included 30,124 head CTs in a non-interpretative paper assessing turn-around times [59]. Test sets used in studies of reader performance ranged in size from 300 to 2568 patients [36, 45, 53]. Two papers compared algorithm and unassisted reader diagnostic accuracy [45, 53]. One paper assessed agreement between manual and automated volumetric measurements [48], and 3 papers compared performance in assisted versus unassisted readers with fully crossed multi-reader multi-case (MRMC) design and heterogeneous multicenter test data [36, 54, 55]. One of these was registered as a clinical trial on clinicaltrials.gov [54]. There was one prospective randomized controlled trial, which employed random notification drop out to assess effects on reporting turnaround time [47]. Twenty papers used retrospective design.

Five papers (24%) compared performance in subgroups for hidden bias/stratification [46, 51, 54–56]. Two papers assessed algorithm performance exclusively in a pediatric cohort [53, 56].

Effects on turnaround time were assessed in 3 papers [42, 47, 59], and one of these [59] evaluated length of stay as an outcome using cohorts before and after product deployment.

Nine of 21 studies (43%) reported industry funding, one reported indirect funding through an NIH T32 mechanism, and one reported funding through a non-profit foundation. Twelve works were conducted in the USA, 1 in India, 2 in Australia, 1 in Switzerland, and 5 were conducted in EU member countries.

CAD prototypes

In total, 174 of 212 (82%) of papers reporting development or implementation of CAD prototype tools employed convolutional neural networks (CNNs). There were no works using more recent deep learning implementations such as vision transformers at time of last search. All papers employing CNNs were published between 2016 and 2022, with the first use case of CNNs for detection of posterior element fractures in the spine published by Roth et al. [60]. This is the only trauma-related CNN-based approach published in 2016 using our search criteria. The work combined feature-engineering using statistical atlases for segmentation with a 2.5D CNN-based approach for classification of segmented regions. This was followed by five deep learning (DL)–based prototype CAD publications in 2017, nine in 2018, 19 in 2019, 44 in 2020, and 69 in 2021.

Unsupervised hand-crafted feature engineering approaches were employed in 33 papers, within a steady range of 2–7 papers per year. Of these papers, 25/33 (76%) described development of novel or pseudo-novel approaches, compared with only 76/174 (44%) of papers using DL-based methods. The other 56% of DL papers employed out-of-the-box implementations with no modification. Novelty was largely employed to optimize detection tasks.

Regarding advanced methods for augmenting dataset size, throughput, or labeling, stressed in the NIH/RSNA/ACR/Academy foundational roadmap [33], only one such paper described a distributed privacy preserving method akin to federated learning [61]; two papers addressed data scarcity using synthetic data augmentation with generative models [62, 63], and two papers used semi-supervised approaches to generate pseudo-labels in unlabeled data [63, 64]. Three papers described human-in-the-loop (HITL) active learning strategies to accelerate labeling and improve accuracy [65–67]. Three foundational papers proposed novel active learning approaches that simulate HITL interaction using committee models in place of multi-reader arbitration [68], easy-to-hard curriculum learning incorporating expert knowledge-based confidence scores [69], and a structural causal model that mimics human counterfactual input by mining object detection results for potential cooccurring fractures that could influence AO fracture grade assignment [70]. No paper employed recently developed out-of-the-box HITL active learning methods for segmentation tasks such as MONAI Label or DeepEdit [71, 72].

All but 4 (11%) of hand-crafted feature engineering papers were noted to have explainable methods, whereas 52 of 174 (30%) of papers with CNN-based methods were entirely black box. Feature engineering papers often involved intrinsic interpretability with techniques that simulated human reasoning for a given task. Examples included fracture-dislocation complexity grading based on number, shape, and size of fracture fragments, and measurements of angulation or displacement in fracture-dislocation [73–75]. While supervised machine learning is much more robust to unseen cases, learned features are less transparent. The three most common methods of post hoc explainability for CNN-based methods included activation maps in 49 papers (40%), segmentations in 27 (22%), and box detections in 21 (17%). Several papers described automatic caliper measurements for vertebral compression [76, 77], one paper described cerebral midline shift measurement [78], and one paper combined CNNs and feature engineering for calcaneal fracture landmark localization and Bohler and Gissane angle measurements [79].

The procedure for determining ground truth was not mentioned in 35 of 212 CAD prototype papers (17%); independent reads with arbitration were used in 25 studies (12%); consensus reads were used in 29 studies (14%); and a single reader determined ground truth in 86 studies (41%). Reports or the EMR were employed in 35 studies (17%). Measurements of reader agreement, interobserver variability, or repeatability were provided in only 13 studies (6%).

Only 15 of 212 studies used public datasets (7%). Siloed datasets from more than one institution were employed in another 42 studies (20%). In total, 155 studies (73%) employed single-center data. Sources of public data included the MICCAI RibFrac CT rib fracture challenge (660 CT studies) [65], the Manitoba bone mineral density registry (~ 12,000 DEXA studies) [80], the MURA musculoskeletal radiograph dataset (~ 14,000 plain radiograph studies) [81], the SIIM-ACR pneumothorax dataset (~ 12,000 plain radiography studies) [82], the RSNA intracranial hemorrhage dataset (~ 25,000 CT studies) [83], and the NIH ChestX-Ray14 dataset (~ 30,000 unique patients with over 100,000 studies) [84].

Thirty-four studies (17%) benchmarked AI performance to unassisted radiologist performance, and 20 studies (9%) compared reader performance with and without AI assistance, with 10 of these studies employing a fully crossed MRMC study design. Of these 34 papers, only 2 evaluated for hidden bias/stratification [85, 86]. No CAD prototype diagnostic performance study used prospective methodology. Fifteen of the 212 CAD prototype studies (7%) involved pediatric populations. There were no user acceptance studies, and no study included an interactive HITL component that affected algorithm output, although one study allowed manual clicks to identify true positive bounding boxes and suggested future incorporation of interactive case–based reasoning [87].

Seventeen studies (8%) evaluated a variety of short- and long-term patient outcomes. These included future risk of non-vertebral fractures from DEXA scan [88]; association of pneumothorax CT volumes and decision to perform chest tube drainage [89]; risk of hemorrhage related complications from traumatic pelvic hematoma [39], hemoperitoneum [90], pelvic fractures [91], and liver laceration [92]; prediction of spinal cord injury clinical severity and motor impairment [93, 94]; association between TBI and discharge Glasgow Coma Scale score [95], concussion severity [96], cognitive impairment [97], Glasgow Outcome Scale score [78, 98, 99], and 6-month mortality [100]; and correlation of pulmonary contusion with risk of mechanical ventilation, increased hospital length of stay [101], and ARDS [102]. Eleven of these 17 papers (64%) involved segmentation and quantitative imaging.

The top five global sources for foundational or clinical translational research productivity were the USA (63 papers), China (36 papers), the EU (23 papers), South Korea (19 papers), and Japan (15 papers). A total of 102 CAD prototype papers (48%) reported research grant funding from a government agency; 20 papers (9%) reported institution, society, or foundation grants; and 11 papers (5%) disclosed industry support. Seventy-four percent of papers from South Korea (14/19), 64% from China (23/36), 54% from the US (34/63), 43% from the EU (10/23), and 7% from Japan (1/15) reported government agency research support. Twenty-six papers reported NIH funding; of these, 11 were in MSK, 6 in neuroimaging, 7 in abdomen/pelvis, and 2 in cardiovascular/chest domains. Overall research output by modality, body region, and year mirrored the growth in studies using FDA-approved tools (Fig. 2).

Discussion

In this work, we aimed to summarize the current state of evolution and maturity of trauma AI/ML CAD tools, clarify key gaps in their technological development, and identify understudied areas that may help anticipate clinical trends and guide future research priorities in the trauma imaging subfield of emergency radiology. To the best of our knowledge, this is the first structured review paper in this domain. Included papers spanned the literature from early proof of concept to validation of commercialized tools.

Traumatic injury is a long-tailed problem with many injury types of varying incidences for any given body region or system. Trends reflected in papers describing prototype tools presaged trends in commercialization. The first paper meeting inclusion criteria that described DL-based methods was published in 2016 [60], and all commercial products, the first described in 2018 [38], employed CNNs. Most prototype and commercial tools focused on detection tasks in the MSK and neuroimaging domains and employed activation maps or box detections for explainability. Although public data were used in only 7% of prototype studies, 7 of the 10 commercial tools were in domains where large public datasets are available, suggesting that public data may be a major driver of late-stage R&D. These included datasets for ICH detection on CT, fracture detection on MSK radiographs, and fracture or pneumothorax detection on chest radiographs.

Approximately 84% of studies described siloed datasets with fewer than 5000 patients. Cross-sectional imaging datasets for abdominopelvic and chest trauma ranged from fewer than 100 to 778 patients, and no commercial products were described in these domains. Torso pathology including organ injury, contusion, and hemorrhage is highly variable in size and appearance with small target to volume ratios, and multiscale DL-based tools for torso pathology have been late-comers [103] and were not reported for trauma until 2020 [104].

There is a need for improved reporting of methodology for ground truth interpretations, higher quality validation data, and outcomes research using commercial CAD tools. Fully crossed MRMC study design with heterogeneous multi-center data was employed in 3 of 21 papers evaluating commercial products, and one paper reported prospective randomized controlled study design for a turnaround time endpoint. Davis et al. evaluated length of hospital stay as an endpoint [59], and this was the only work using a commercial product to assess patient outcomes. Fewer than one quarter of commercial CAD papers performed subgroup analysis for potential sources of bias. High-quality ground truth based on independent reads and a method of arbitration was described in 9 of 21 papers, and only one paper reported the level of agreement between experts. One paper reported registration in clinicaltrials.gov [54].

There is an ongoing need for performance and outcomes studies following regulatory approval. As studies accumulate, data can be aggregated in meta-analyses to produce a higher level of evidence to justify institutional adoption by academic and private practice stakeholders involved in radiology AI governance.

Seventy-six of papers described hand-crafted feature engineering methods, but only 44% of papers using supervised DL-based approaches described some level of novelty. This may speak to the versatility and democratizing effect of supervised machine learning, where existing implementations can be effectively applied to a wide range of pathologies. Of note, 82% of prototype CAD tools employed CNNs, with a steep rise since the first DL-based publication in 2016.

While all commercial tools had some level of explainability, 30% of prototype CAD studies employed black-box approaches. Unsupervised approaches using hand-crafted methods were more commonly inherently interpretable, using measurements or shape-based features that would be employed to arrive at a diagnosis in the course of clinical interpretation. However, these methods were not shown to scale to large data and none led to regulatory approval and commercialization.

No human-in-the-loop CAD methods were described, and no studies assessed parameters of user acceptance. Taken together, with the emergence of scalable DL-based approaches, there is an unmet need for tools that incorporate human–computer interaction. User acceptance should be considered and evaluated early in the R&D process to maximize clinical utility [105]. Standardized methods are needed for subjective or semi-quantitative evaluation of usability.

Overall, there is a strong need for higher quality methodology and reporting to improve transparency and confidence in the data. Researchers should be encouraged to use available checklists such as CLAIM [106] and STARD AI [107, 108]. There were few CAD prototype studies that benchmarked algorithm performance to human performance (17%), and only 6% of these evaluated for hidden bias. This is not surprising given the emphasis on novelty or at least initial proof of concept for any given use case; however, only 6% of all CAD prototype studies reported a high-quality ground truth procedure including reader agreement or repeatability, posing a major limitation with respect to transparency.

Eight percent of CAD prototype studies evaluated patient outcomes, and approximately two thirds of these described segmentation and quantitative visualization tools. However, our findings show that quantitative trauma imaging tools, despite emphasis on outcomes, have not reached the stage of regulatory approval and commercialization. These tools predominate for cross-sectional imaging modalities and typically require painstaking voxel-level ground truth annotation, making dataset curation very challenging. Because such tools provide granular, objective information, their value proposition resides with precision medicine. Therefore, proof of concept feasibility assessments focusing on outcome prediction and method validity in small patient populations is typically considered a prerequisite before scaling to large heterogeneous multicenter datasets.

Few papers described advanced methods that address scarcity of high-quality labeled data. One article described a federated approach, and several used synthetic data augmentation, semi-supervised methods, and active learning strategies. Our search was limited to peer reviewed biomedical literature and technical papers employing these strategies may have been overlooked. However, our review indicates that foundational research efforts to date have not resulted in advanced techniques for augmenting and accelerating data curation that are ready for mainstream adoption.

Forty-eight percent of prototype CAD papers reported research support from government agencies, and 43% of commercial CAD papers reported industry funding. Countries with the highest proportion of funded studies (the USA, China, and South Korea) also had the highest levels of research productivity.

Even though trauma remains the leading cause of death and disability in patients under 45 years of age, trauma imaging remains a relatively small and underfunded branch of radiology. In the field of radiology as a whole, AI/ML publications have increased exponentially, primarily in the fields of neuroradiology, abdominal imaging, and chest imaging, spurred by federal agency and industry-side funding [20]. Our findings suggest that increased funding opportunities, researcher engagement, research training, and institutional buy-in will accelerate research productivity and translation of tools to the trauma setting.

Conclusions

In conclusion, AI CAD tools are likely to improve ER/trauma radiologist productivity and diagnostic performance, reduce turnaround times, decrease ER and hospital stays, and improve survival of severely injured patients. However, these tools are currently in a very early stage of maturity. There are few FDA-approved products for a limited number of use cases, and there has not been sufficient validation of commercial tools to generate meta-analyses. The scarcity of large heterogeneous datasets with high-quality annotation continues to pose a major barrier. There remains an unmet need for out-of-the-box tools that accelerate data labeling and for multicenter privacy-preserving distributed learning.

A greater emphasis should be placed on performance validation data that incorporates assessment of bias and robustness across relevant subgroups. The methodology used for ground truth annotation is highly variable across the body of literature in this area. Researchers should be encouraged to employ independent readers with arbitration and provide data on reader agreement and reproducibility of measurements.

Additionally, the range of techniques for explainability and interpretability using scalable DL-based approaches remains narrow, and methods that build trust through human–computer interaction are lacking. More emphasis should be placed on evaluation of end-user assessment of system benevolence and capability. Finally, an increase in funding opportunities would likely accelerate the R&D pipeline for trauma imaging CAD tools.

A potential future avenue for our expert panel to explore includes a follow-up scoping review in several years to map progress since this review and a position paper focusing on research priorities in ER/trauma imaging. A survey of ASER members' perceptions and expectations with respect to AI/ML CAD tools is also forthcoming.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Funding

David Dreizin funding source: NIH K08 EB027141-01A1 (PI: David Dreizin, MD)

References

1. Castro DC, Walker I, Glocker B (2020) Causality matters in medical imaging. Nat Commun 11(1):1–10 [PubMed: 31911652]
2. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, Folio LR, Summers RM, Rubin DL, Lungren MP (2020) Preparing medical imaging data for machine learning. Radiology 295(1):4–15 [PubMed: 32068507]
3. Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D (2017) Interpretive error in radiology. Am J Roentgenol 208(4):739–749 [PubMed: 28026210]

4. Chung JH, Strigel RM, Chew AR, Albrecht E, Gunn ML (2009) Overnight resident interpretation of torso CT at a level 1 trauma center: an analysis and review of the literature. Acad Radiol 16(9):1155–1160 [PubMed: 19481962]

5. Bruno MA, Duncan JR, Bierhals AJ, Tappouni R (2018) Over-night resident versus 24-hour attending radiologist coverage in academic medical centers. Radiology 289(3):809–813 [PubMed: 30277849]

6. Banaste N, Caurier B, Bratan F, Bergerot J-F, Thomson V, Millet I (2018) Whole-body CT in patients with multiple traumas: factors leading to missed injury. Radiology 289(2):374–383 [PubMed: 30084754]

7. Glover M IV, Almeida RR, Schaefer PW, Lev MH, Mehan WA Jr (2017) Quantifying the impact of noninterpretive tasks on radiology report turn-around times. J Am Coll Radiol 14(11):1498–1503 [PubMed: 28916177]

8. Hunter TB, Taljanovic MS, Krupinski E, Ovitt T, Stubbs AY (2007) Academic radiologists' on-call and late-evening duties. J Am Coll Radiol 4(10):716–719 [PubMed: 17903757]

9. Hanna TN, Loehfelm T, Khosa F, Rohatgi S, Johnson J-O (2016) Overnight shift work: factors contributing to diagnostic discrepancies. Emerg Radiol 23(1):41–47 [PubMed: 26475281]

10. Barquist ES, Pizano LR, Feuer W, Pappas PA, McKenney KA, LeBlang SD, Henry RP, Rivas LA, Cohn SM (2004) Inter-and intrarater reliability in computed axial tomographic grading of splenic injury: why so many grading scales? J Trauma Acute Care Surg 56(2):334–338

11. Clark R, Hird K, Misur P, Ramsay D, Mendelson R (2011) CT grading scales for splenic injury: why can't we agree? J Med Imaging Radiat Oncol 55(2):163–169 [PubMed: 21501405]

12. Chen H, Unberath M, Dreizin D (2023) Toward automated interpretable AAST grading for blunt splenic injury. Emerg Radiol 30(1):41–50. 10.1007/s10140-022-02099-1 [PubMed: 36371579]

13. Furey AJ, O'Toole RV, Nascone JW, Sciadini MF, Copeland CE, Turen C (2009) Classification of pelvic fractures: analysis of inter-and intraobserver variability using the Young-Burgess and Tile classification systems. Orthopedics (Online) 32(6):401

14. Liu J, Varghese B, Taravat F, Eibschutz LS, Gholamrezanezhad A (2022) An extra set of intelligent eyes: application of artificial intelligence in imaging of abdominopelvic pathologies in emergency radiology. Diagnostics 12(6):1351 [PubMed: 35741161]

15. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition p. 770–778

17. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A (eds) Medical image computing and computer-assisted intervention – MICCAI 2015. MICCAI 2015. Lecture notes in computer science(), vol 9351. Springer, Cham. 10.1007/978-3-319-24574-4_28

18. Zhou SK, Greenspan H, Davatzikos C, Duncan JS, Van Gin-neken B, Madabhushi A, Prince JL, Rueckert D, Summers RM (2021) A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. Proc IEEE 109(5):820–838

19. Fujita H (2020) AI-based computer-aided diagnosis (AI-CAD): the latest review to read first. Radiol Phys Technol 13(1):6–19 [PubMed: 31898014]

20. West E, Mutasa S, Zhu Z, Ha R (2019) Global trend in artificial intelligence–based publications in radiology from 2000 to 2018. Am J Roentgenol 213(6):1204–1206 [PubMed: 31414886]

21. Harvey HB, Gowda V (2020) How the FDA regulates AI. Acad Radiol 27(1):58–61 [PubMed: 31818387]

22. Ebrahimian S, Kalra MK, Agarwal S, Bizzo BC, Elkholy M, Wald C, Allen B, Dreyer KJ (2022) FDA-regulated AI algorithms: trends, strengths, and gaps of validation studies. Acad Radiol 29(4):559–566 [PubMed: 34969610]

23. Sammer MB, Sher AC, Towbin AJ (2022) Ensuring adequate development and appropriate use of artificial intelligence in pediatric medical imaging. Am J Roentgenol 218(1):182–183 [PubMed: 34319165]

24. Yang L, Ene IC, Arabi Belaghi R, Koff D, Stein N, Santaguida PL (2022) Stakeholders' perspectives on the future of artificial intelligence in radiology: a scoping review. Eur Radiol 32(3):1477–1495. 10.1007/s00330-021-08214-z [PubMed: 34545445]

25. Benjamens S, Dhunnoo P, Meskó B (2020) The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med 3(1):1–8 [PubMed: 31934645]

26. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K (2019) The practical implementation of artificial intelligence technologies in medicine. Nat Med 25(1):30–36 [PubMed: 30617336]

27. Dreizin D, Munera F (2012) Blunt polytrauma: evaluation with 64-section whole-body CT angiography. Radiographics 32(3):609–631. 10.1148/rg.323115099 [PubMed: 22582350]

28. Dreizin D, Munera F (2015) Multidetector CT for penetrating torso trauma: state of the art. Radiology 277(2):338–355 [PubMed: 26492022]

29. Varoquaux G, Cheplygina V (2022) Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ Digit Med 5(1):1–8 [PubMed: 35013539]

30. Weikert T, Cyriac J, Yang S, Nesic I, Parmar V, Stieltjes B (2020) A practical guide to artificial intelligence–based image analysis in radiology. Invest Radiol 55(1):1–7 [PubMed: 31503083]

31. Arksey H, O'Malley L (2005) Scoping studies: towards a methodological framework. Int J Soc Res Methodol 8(1):19–32

32. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA (2014) A scoping review of scoping reviews: advancing the approach and enhancing the consistency. Res Synth Meth 5(4):371–385

33. Langlotz CP, Allen B, Erickson BJ, Kalpathy-Cramer J, Bigelow K, Cook TS, Flanders AE, Lungren MP, Mendelson DS, Rudie JD (2019) A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology 291(3):781 [PubMed: 30990384]

34. Allen B Jr, Seltzer SE, Langlotz CP, Dreyer KP, Summers RM, Petrick N, Marinac-Dabic D, Cruz M, Alkasab TK, Hanisch RJ (2019) A road map for translational research on artificial intelligence in medical imaging: from the 2018 National Institutes of Health/RSNA/ACR/The Academy Workshop. J Am Coll Radiol 16(9):1179–1189 [PubMed: 31151893]

35. Majkowska A, Mittal S, Steiner DF, Reicher JJ, McKinney SM, Duggan GE, Eswaran K, Cameron Chen P-H, Liu Y, Kalidindi SR (2020) Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. Radiology 294(2):421–431 [PubMed: 31793848]

36. Seah JC, Tang CH, Buchlak QD, Holt XG, Wardman JB, Aimoldin A, Esmaili N, Ahmad H, Pham H, Lambert JF (2021) Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. Lancet Digit Health 3(8):e496–e506 [PubMed: 34219054]

37. Jones RM, Sharma A, Hotchkiss R, Sperling JW, Hamburger J, Ledig C, O'Toole R, Gardner M, Venkatesh S, Roberts MM (2020) Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. NPJ Digit Med 3(1):1–6 [PubMed: 31934645]

38. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, Mahajan V, Rao P, Warier P (2018) Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 392(10162):2388–2396 [PubMed: 30318264]

39. Dreizin D, Zhou Y, Chen T, Li G, Yuille AL, McLenithan A, Morrison JJ (2020) Deep learning-based quantitative visualization and measurement of extraperitoneal hematoma volumes in patients with pelvic fractures: potential role in personalized forecasting and decision support. J Trauma Acute Care Surg 88(3):425 [PubMed: 32107356]

40. Harris RJ, Kim S, Lohr J, Towey S, Velichkovich Z, Kabachenko T, Driscoll I, Baker B (2019) Classification of aortic dissection and rupture on post-contrast CT images using a convolutional neural network. J Digit Imaging 32(6):939–946 [PubMed: 31515752]

41. Ginat DT (2020) Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage. Neuroradiology 62(3):335–340 [PubMed: 31828361]

42. Ginat D (2021) Implementation of machine learning software on the radiology worklist decreases scan view delay for the detection of intracranial hemorrhage on CT. Brain Sci 11(7):832 [PubMed: 34201775]

43. Kundisch A, Hönning A, Mutze S, Kreissl L, Spohn F, Lemcke J, Sitz M, Sparenberg P, Goelz L (2021) Deep learning algorithm in detecting intracranial hemorrhages on emergency computed tomographies. PLoS ONE 16(11):e0260560 [PubMed: 34843559]
44. Ojeda P, Zawaideh M, Mossa-Basha M, Haynor D (2019) The ional neural network for detection of intracranial bleeds on non-contrast head computed tomography studies. In: Proc. SPIE 10949, Medical Imaging 2019: Image processing, 109493J. 10.1117/12.2513167
45. Kau T, Ziurlys M, Taschwer M, Kloss-Brandstätter A, Grabner G, Deutschmann H (2022) FDA-approved deep learning software application versus radiologists with different levels of expertise: detection of intracranial hemorrhage in a retrospective single-center study. Neuroradiology 64(5):981–990 [PubMed: 34988593]
46. Voter AF, Meram E, Garrett JW, John-Paul JY (2021) Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of intracranial hemorrhage. J Am Coll Radiol 18(8):1143–1152 [PubMed: 33819478]
47. Wismüller A, Stockmaster L (2020) A prospective randomized clinical trial for measuring radiology study reporting time on artificial intelligence-based detection of intracranial hemorrhage in emergent care head CT. In: Proc. SPIE 11317, Medical Imaging 2020: Biomedical applications in molecular, structural, and functional imaging, 113170M. 10.1117/12.2552400
48. Heit J, Coelho H, Lima F, Granja M, Aghaebrahim A, Hanel R, Kwok K, Haerian H, Cereda C, Venkatasubramanian C (2021) Automated cerebral hemorrhage detection using RAPID. Am J Neuroradiol 42(2):273–278 [PubMed: 33361378]
49. Gipson J, Tang V, Seah J, Kavnoudias H, Zia A, Lee R, Mitra B, Clements W (2022) Diagnostic accuracy of a commercially available deep-learning algorithm in supine chest radiographs following trauma. Br J Radiol 95:20210979 [PubMed: 35271382]
50. Small J, Osler P, Paul A, Kunst M (2021) Ct cervical spine fracture detection using a convolutional neural network. Am J Neuroradiol 42(7):1341–1347 [PubMed: 34255730]
51. Voter A, Larson M, Garrett J, Yu J-P (2021) Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of cervical spine fractures. Am J Neuroradiol 42(8):1550–1556 [PubMed: 34117018]
52. Weikert T, Noordtzij LA, Bremerich J, Stieltjes B, Parmar V, Cyriac J, Sommer G, Sauter AW (2020) Assessment of a deep learning algorithm for the detection of rib fractures on whole-body trauma computed tomography. Korean J Radiol 21(7):891 [PubMed: 32524789]
53. Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard N-E, Guermazi A (n.d.) Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skelet Radiol 2022:1–11
54. Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard NE, Guermazi A (2022) Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skeletal Radiol 51(11):2129–2139. 10.1007/s00256-022-04070-0 [PubMed: 35522332]
55. Duron L, Ducarouge A, Gillibert A, Lainé J, Allouche C, Cherel N, Zhang Z, Nitche N, Lacave E, Pourchot A (2021) Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. Radiology 300(1):120–129 [PubMed: 33944629]
56. Dupuis M, Delbos L, Veil R, Adamsbaum C (2022) External validation of a commercially available deep learning algorithm for fracture detection in children. Diagn Interv Imaging 103(3):151–159 [PubMed: 34810137]
57. Rueckel J, Sperl JI, Kaestle S, Hoppe BF, Fink N, Rudolph J, Schwarze V, Geyer T, Strobl FF, Ricke J (2021) Reduction of missed thoracic findings in emergency whole-body computed tomography using artificial intelligence assistance. Quant Imaging Med Surg 11:2486–2498 [PubMed: 34079718]
58. Genant HK, Li J, Wu CY, Shepherd JA (2000) Vertebral fractures in osteoporosis: a new method for clinical assessment. J Clin Densitom 3(3):281–290 [PubMed: 11090235]
59. Davis MA, Rao B, Cedeno PA, Saha A, Zohrabian VM (2022) Machine learning and improved quality metrics in acute intracranial hemorrhage by noncontrast computed tomography. Curr Probl Diagn Radiol 51(4):556–561. 10.1067/j.cpradiol.2020.10.007 [PubMed: 33243455]

60. Shin H-C et al. (2016) Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging 35(5):1285–1298. 10.1109/TMI.2016.2528162 [PubMed: 26886976]
61. Remedios SW, Roy S, Bermudez C, Patel MB, Butman JA, Landman BA, Pham DL (2020) Distributed deep learning across multisite datasets for generalized CT hemorrhage segmentation. Med Phys 47(1):89–98 [PubMed: 31660621]
62. Mutasa S, Varada S, Goel A, Wong TT, Rasiej MJ (2020) Advanced deep learning techniques applied to automated femoral neck fracture detection and classification. J Digit Imaging 33(5):1209–1217 [PubMed: 32583277]
63. Zhou Y, Dreizin D, Wang Y, Liu F, Shen W, Yuille AL (2021) External attention assisted multi-phase splenic vascular injury segmentation with limited data. IEEE Trans Med Imaging 41(6):1346–1357
64. Lind A, Akbarian E, Olsson S, Nåsell H, Sköldenberg O, Razavian AS, Gordon M (2021) Artificial intelligence for the classification of fractures around the knee in adults according to the 2018 AO/OTA classification system. PLoS ONE 16(4):e0248809 [PubMed: 33793601]
65. Jin L, Yang J, Kuang K, Ni B, Gao Y, Sun Y, Gao P, Ma W, Tan M, Kang H (2020) Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet. EBioMedicine 62:103106 [PubMed: 33186809]
66. Zhou Q-Q, Hu Z-C, Tang W, Xia Z-Y, Wang J, Zhang R, Li X, Chen C-Y, Zhang B, Lu L (2022) Precise anatomical localization and classification of rib fractures on CT using a convolutional neural network. Clin Imaging 81:24–32 [PubMed: 34598000]
67. Olczak J, Emilson F, Razavian A, Antonsson T, Stark A, Gordon M (2020) Ankle fracture classification using deep learning: automating detailed AO Foundation/Orthopedic Trauma Association (AO/OTA) 2018 malleolar fracture identification reaches a high degree of correct classification. Acta Orthop 92(1):102–108 [PubMed: 33103536]
68. Huang Y-J, Liu W, Wang X, Fang Q, Wang R, Wang Y, Chen H, Chen H, Meng D, Wang L (2020) Rectifying supporting regions with mixed and active supervision for rib fracture recognition. IEEE Trans Med Imaging 39(12):3843–3854 [PubMed: 32746128]
69. Luo J, Kitamura G, Doganay E, Arefan D, Wu S (2021) Medical knowledge-guided deep curriculum learning for elbow fracture diagnosis from x-ray images. In: Proc. SPIE 11597, Medical Imaging 2021: Computer-aided diagnosis, 1159712. 10.1117/12.2582184
70. Zapaishchykova A, Dreizin D, Li Z, Wu JY, Faghihroohi S, Unberath M (2021) An interpretable approach to automated severity scoring in pelvic trauma. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021, vol 12903. Lecture Notes in Computer Science(), Springer, Cham. 10.1007/978-3-030-87199-4_40
71. Diaz-Pinto A, Alle S, Ihsani A, Asad M, Nath V, Pérez-García F, Mehta P, Li W, Roth HR, Vercauteren T (2022) Monai label: a framework for ai-assisted interactive labeling of 3d medical images. arXiv preprint arXiv:220312362
72. Diaz-Pinto A, Mehta P, Alle S, Asad M, Brown R, Nath V, Ihsani A, Antonelli M, Palkovics D, Pinter C (2022) DeepEdit: deep editable learning for interactive segmentation of 3D medical images. MICCAI Workshop on Data Augmentation, Labelling, and Imperfections: Springer, p. 11–21
73. Burns JE, Yao J, Muñoz H, Summers RM (2016) Automated detection, localization, and classification of traumatic vertebral body fractures in the thoracic and lumbar spine at CT. Radiology 278(1):64 [PubMed: 26172532]
74. Bandyopadhyay O, Biswas A, Bhattacharya BB (2016) Long-bone fracture detection in digital X-ray images based on digital-geometric techniques. Comput Methods Programs Biomed 123:2–14 [PubMed: 26477855]
75. Sun L, Kong Q, Huang Y, Yang J, Wang S, Zou R, Yin Y, Peng J (2020) Automatic segmentation and measurement on knee computerized tomography images for patellar dislocation diagnosis. Comput Math Methods Med 2020
76. Seo JW, Lim SH, Jeong JG, Kim YJ, Kim KG, Jeon JY (2021) A deep learning algorithm for automated measurement of vertebral body compression from X-ray images. Sci Rep 11(1):1–10 [PubMed: 33414495]

77. Baum T, Bauer JS, Klinder T, Dobritz M, Rummeny EJ, Noël PB, Lorenz C (2014) Automatic detection of osteoporotic vertebral fractures in routine thoracic and abdominal MDCT. Eur Radiol 24(4):872–880 [PubMed: 24425527]
78. Xia X, Zhang X, Huang Z, Ren Q, Li H, Li Y, Liang K, Wang H, Han K, Meng X (2021) Automated detection of 3D midline shift in spontaneous supratentorial intracerebral haemorrhage with non-contrast computed tomography using deep convolutional neural networks. Am J Transl Res 13(10):11513 [PubMed: 34786077]
79. Guo J, Mu Y, Xue D, Li H, Chen J, Yan H, Xu H, Wang W (2021) Automatic analysis system of calcaneus radiograph: Rotation-invariant landmark detection for calcaneal angle measurement, fracture identification and fracture region segmentation. Comput Methods Programs Biomed 206:106124 [PubMed: 34004502]
80. Monchka BA, Kimelman D, Lix LM, Leslie WD (2021) Feasibility of a generalized convolutional neural network for automated identification of vertebral compression fractures: the Manitoba Bone Mineral Density Registry. Bone 150:116017 [PubMed: 34020078]
81. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, Yang B, Zhu K, Laird D, Ball RL (2017) Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:171206957
82. Wang Y, Wang K, Peng X, Shi L, Sun J, Zheng S, Shan F, Shi W, Liu L (2021) DeepSDM: Boundary-aware pneumothorax segmentation in chest X-ray images. Neurocomputing 454:201–211
83. Flanders AE, Prevedello LM, Shih G, Halabi SS, Kalpathy-Cramer J, Ball R, Mongan JT, Stein A, Kitamura FC, Lungren MP (2020) Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge. Radiology: Artif Intell 2(3):e190211
84. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) Chestx-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. Proceedings of the IEEE conference on computer vision and pattern recognition, p. 2097–2106
85. Choi JW, Cho YJ, Ha JY, Lee YY, Koh SY, Seo JY, Choi YH, Cheon J-E, Phi JH, Kim I (2022) Deep learning-assisted diagnosis of pediatric skull fractures on plain radiographs. Korean J Radiol 23(3):343 [PubMed: 35029078]
86. Oakden-Rayner L, Gale W, Bonham TA, Lungren MP, Carneiro G, Bradley AP, Palmer LJ (2022) Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet Digit Health 4(5):e351–e358 [PubMed: 35396184]
87. Jiménez-Sánchez A, Kazi A, Albarqouni S, Kirchhoff C, Biberthaler P, Navab N, Kirchhoff S, Mateus D (2020) Precise proximal femur fracture classification for interactive training and surgical planning. Int J Comput Assist Radiol Surg 15(5):847–857 [PubMed: 32335786]
88. Derkatch S, Kirby C, Kimelman D, Jozani MJ, Davidson JM, Leslie WD (2019) Identification of vertebral fractures by convolutional neural networks to predict nonvertebral and hip fractures: a registry-based cohort study of dual X-ray absorptiometry. Radiology 293(2):405–411 [PubMed: 31526255]
89. Cai W, Lee J-G, Fikry K, Yoshida H, Novelline R, de Moya M (2012) MDCT quantification is the dominant parameter in decision-making regarding chest tube drainage for stable patients with traumatic pneumothorax. Comput Med Imaging Graph 36(5):375–386 [PubMed: 22560899]
90. Dreizin D, Zhou Y, Fu S, Wang Y, Li G, Champ K, Siegel E, Wang Z, Chen T, Yuille AL (2020) A multiscale deep learning method for quantitative visualization of traumatic hemoperitoneum at CT: Assessment of feasibility and comparison with subjective categorical estimation. Radiol Artif Intell 2(6):e190220. 10.1148/ryai.2020190220 [PubMed: 33330848]
91. Dreizin D, Goldmann F, LeBedis C, Boscak A, Dattwyler M, Bodanapally U, Li G, Anderson S, Maier A, Unberath M (2021) An automated deep learning method for tile AO/OTA pelvic fracture severity grading from trauma whole-body CT. J Digit Imaging 34(1):53–65 [PubMed: 33479859]
92. Dreizin D, Chen T, Liang Y, Zhou Y, Paes F, Wang Y, Yuille AL, Roth P, Champ K, Li G (2021) Added value of deep learning-based liver parenchymal CT volumetry for predicting major arterial injury after blunt hepatic trauma: a decision tree analysis. Abdom Radiol 46(6):2556–2566

93. Okimatsu S, Maki S, Furuya T, Fujiyoshi T, Kitamura M, Inada T, Aramomi M, Yamauchi T, Miyamoto T, Inoue T (2022) Determining the short-term neurological prognosis for acute cervical spinal cord injury using machine learning. J Clin Neurosci 96:74–79 [PubMed: 34998207]

94. McCoy D, Dupont S, Gros C, Cohen-Adad J, Huie R, Ferguson A, Duong-Fernandez X, Thomas L, Singh V, Narvid J (2019) Convolutional neural network–based automated segmentation of the spinal cord and contusion injury: deep learning biomarker correlates of motor impairment in acute spinal cord injury. Am J Neuroradiol 40(4):737–744 [PubMed: 30923086]

95. Chaganti S, Plassard AJ, Wilson L, Smith MA, Patel MB, Landman BA (2016) A Bayesian framework for early risk prediction in traumatic brain injury. Proc SPIE Int Soc Opt Eng 27(9784):978422. 10.1117/12.2217306

96. Cai Y, Wu S, Zhao W, Li Z, Wu Z, Ji S (2018) Concussion classification via deep learning using whole-brain white matter fiber strains. PLoS ONE 13(5):e0197992 [PubMed: 29795640]

97. Hellyer PJ, Leech R, Ham TE, Bonnelle V, Sharp DJ (2013) Individual prediction of white matter injury following traumatic brain injury. Ann Neurol 73(4):489–499 [PubMed: 23426980]

98. Kim Y-T, Kim H, Lee C-H, Yoon BC, Kim JB, Choi YH, Cho W-S, Oh B-M, Kim D-J (2021) Intracranial densitometry-augmented machine learning enhances the prognostic value of brain CT in pediatric patients with traumatic brain injury: A retrospective pilot study. Front Pediatr 9:750272. 10.3389/fped.2021.750272 [PubMed: 34796154]

99. Mohamed M, Alamri A, Mohamed M, Khalid N, O'Halloran P, Staartjes V, Uff C (2022) Prognosticating outcome using magnetic resonance imaging in patients with moderate to severe traumatic brain injury: A machine learning approach. Brain Inj 36(3):353–358. 10.1080/02699052.2022.203418 [PubMed: 35129403]

100. Yao H, Williamson C, Gryak J, Najarian K (2020) Automated hematoma segmentation and outcome prediction for patients with traumatic brain injury. Artif Intell Med 107:101910 [PubMed: 32828449]

101. Choi J, Mavrommati K, Li NY, Patil A, Chen K, Hindin DI, For-rester JD (2022) Scalable deep learning algorithm to compute percent pulmonary contusion among patients with rib fractures. J Trauma Acute Care Surg 93(4):461–466 [PubMed: 35319542]

102. Röhrich S, Hofmanninger J, Negrin L, Langs G, Prosch H (2021) Radiomics score predicts acute respiratory distress syndrome based on the initial CT scan after trauma. Eur Radiol 31(8):5443–5453 [PubMed: 33733689]

103. Lee S, Summers RM (2021) Clinical artificial intelligence applications in radiology: chest and abdomen. Radiol Clin 59(6):987–1002

104. Dreizin D, Zhou Y, Zhang Y, Tirada N, Yuille AL (2020) Performance of a deep learning algorithm for automated segmentation and quantification of traumatic pelvic hematomas on CT. J Digit Imaging 33(1):243–251 [PubMed: 31172331]

105. Chen H, Gomez C, Huang C-M, Unberath M (2022) Explainable medical imaging AI needs human-centered design: guidelines and evidence from a systematic review. npj Digit Med 5(1):1–15 [PubMed: 35013539]

106. Mongan J, Moy L, Kahn CE Jr (2020) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. Radiol Artif Intell 2(2):e200029. 10.1148/ryai.2020200029 [PubMed: 33937821]

107. Sounderajah V, Ashrafian H, Aggarwal R, De Fauw J, Denniston AK, Greaves F, Karthikesalingam A, King D, Liu X, Markar SR (2020) Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI Steering Group. Nat Med 26(6):807–808 [PubMed: 32514173]

108. Sounderajah V, Ashrafian H, Golub RM, Shetty S, De Fauw J, Hooft L, Moons K, Collins G, Moher D, Bossuyt PM (2021) Developing a reporting guideline for artificial intelligence-centred diagnostic test accuracy studies: the STARD-AI protocol. BMJ Open 11(6):e047709

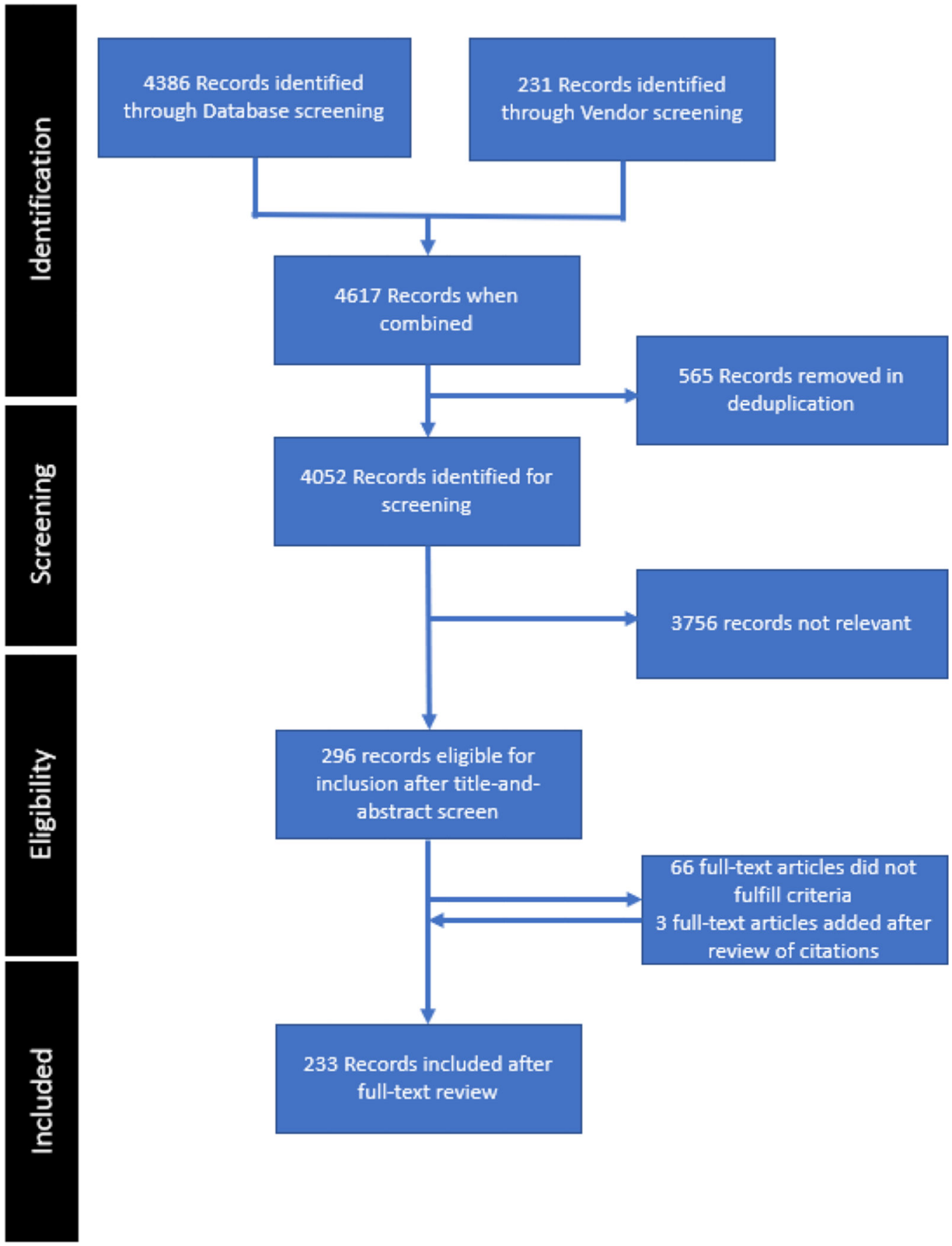

Fig. 1.
Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flowchart depicting study selection

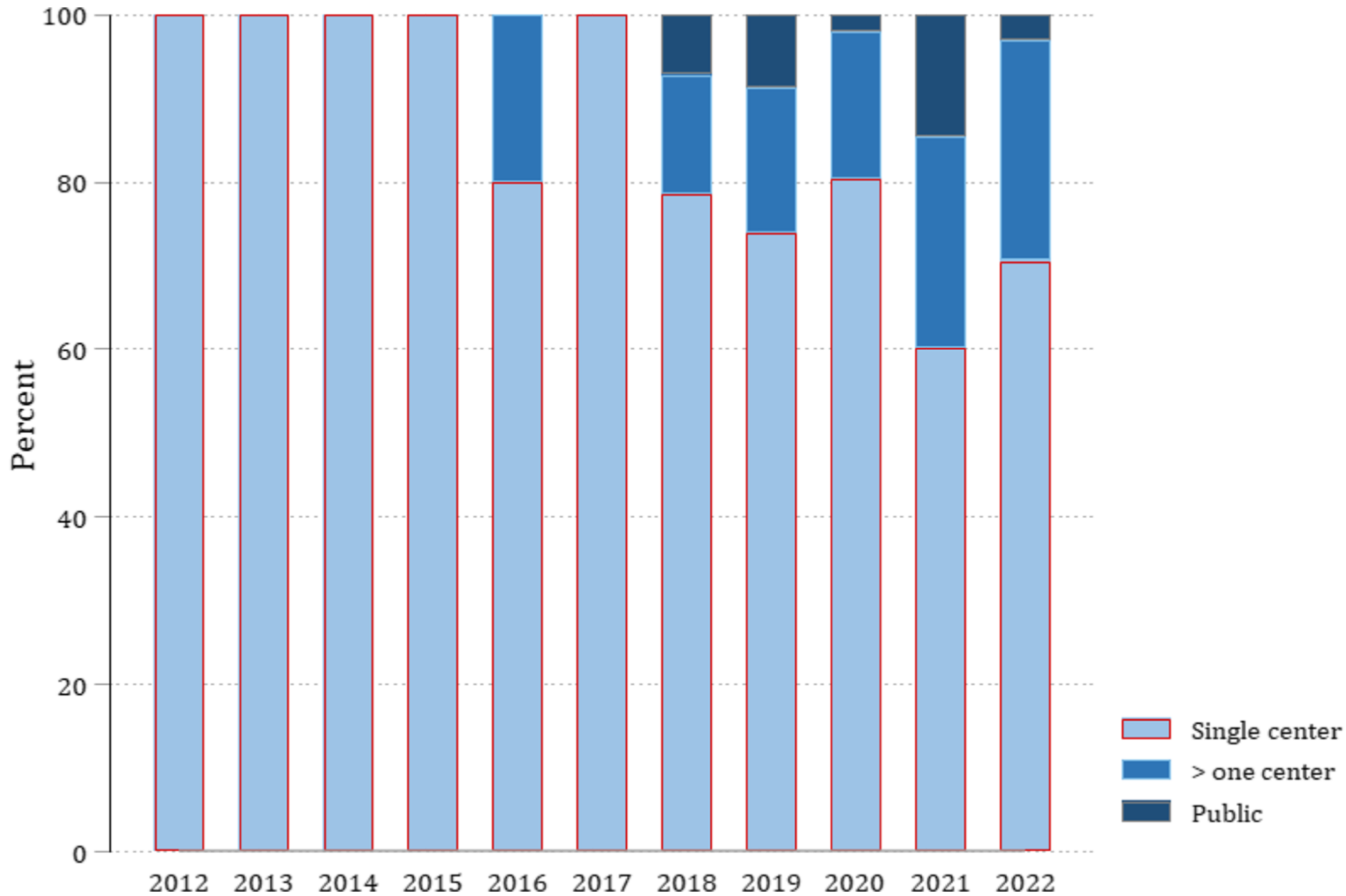

Fig. 2.
Percent plot of publications using siloed (non-public) single center, siloed multicenter, and public data over time

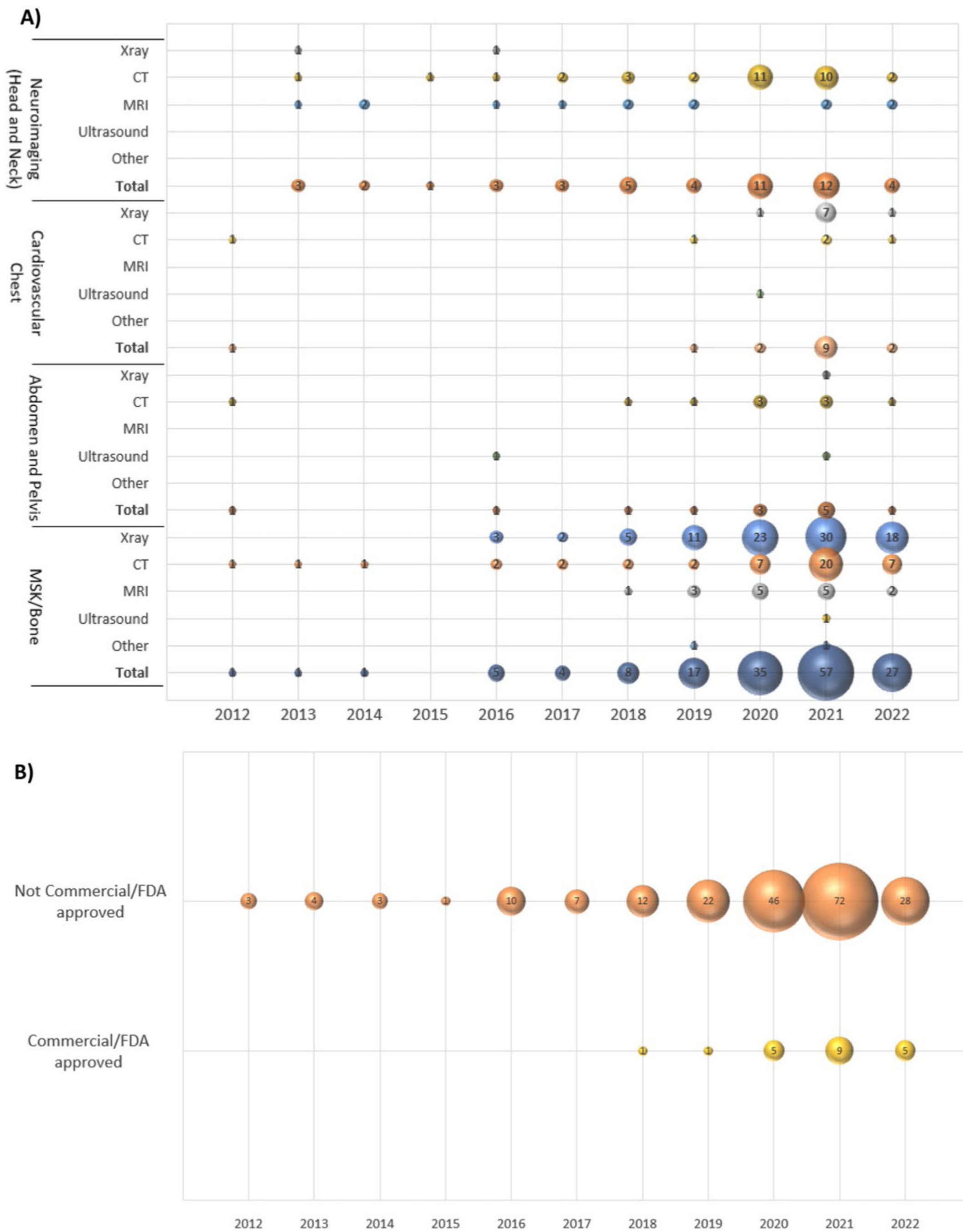

Fig. 3.

Bubble plots depict the number of papers in each domain over time. **A** Papers by body region and imaging modality. **B** Papers by regulatory status. The final search date of June 6, 2022 (midway through year) accounts for fewer publications in 2022

Table 1

General characteristics of the papers included in this review

Characteristic	Number (*n* = 233)	Percentage (%)
Year of publication		
2012–2014	10	4
2015–2016	11	5
2017–2018	21	9
2019–2020	74	32
2021–2022	117	50
Modality		
Radiography	104	45
Computed tomography	94	40
MRI	29	12
Ultrasound	4	2
DEXA	2	1
Body region		
Musculoskeletal/spine	156	67
Head/neck	49	21
Chest/cardiovascular	15	6
Abdomen/pelvis	13	6
Dataset size		
< 100	44	19
100–1000	99	42
1000–5000	45	19
5000–50,000	32	14
50,000–500,000	3	1
> 500,000	2	1
CAD task		
Detection	148	64
Classification	14	6
Detection and classification	27	12
Segmentation/quantitative visualization	31	13
Automated caliper measurements	8	3
Risk stratification/prognostication	5	2

CAD computed-aided detection and diagnosis, *CT* computed tomography, *DEXA* dual X-ray absorptiometry, *MRI* magnetic resonance imaging

Table 2

Study characteristics by regulatory approval

	FDA approved (*n* = 21)	**Non-FDA approved (*n* = 212)**
Type of AI		
CNN	21	174
Novel method	0	76
XAI	21	122
Unsupervised hand-crafted feature engineering	0	33
Novel method	n/a	25
XAI	n/a	29
Types of explainable AI (XAI) in CNN methods	21	122
Activation maps	10	49
Segmentation	3	27
Box detection	5	21
Caliper measurement/other	3	2
Dataset provenance*		
Included public data	1	15
Siloed dataset	20	197
> 1 institution	6	155
Single center	14	42
Advanced approaches to overcome data scarcity		
Federated learning	0	1
Data augmentation	0	2
Pseudo-label generation	0	2
Active learning	0	3
Other	0	3
Ground truth method		
Independent read with arbitration	6	25
Consensus reads	3	29
Single reader	3	86
Other/not mentioned	9	35

	FDA approved (n = 21)	Non-FDA approved (n = 212)
Validation of diagnostic performance		
Algorithm vs unassisted reader diagnostic accuracy	2	34
Automated vs manual volumetric measurement	1	0
Assisted vs unassisted reader diagnostic accuracy	3	20
With fully crossed MRMC design	3	10
Assessed hidden bias/stratification	5	9[†]
Involved pediatric populations	2	15
Evaluated patient/reporter outcomes	3	17
Evaluated user acceptance	0	0
Involved human–computer interaction	0	0
Country of origin [no. with government support]		
US	12 [8]	63 [34]
China	0 [0]	36 [23]
EU	5 [0]	23 [10]
South Korea	0 [0]	19 [14]
Japan	0 [0]	15 [1]
Funding source		
Government agency	1	102
Institution, society, or foundation grant	1	20
Industry support	9	11

CNN convolutional neural networks, *FDA* Food and Drug Administration, *MRMC* multi-reader multi-case, *US* United States, *XAI* explainable artificial intelligence

[*] For papers describing FDA-approved tools, dataset provenance refers to test data

[†] Two of these papers used fully crossed MRMC study design.