European Journal of Radiology 173 (2024) 111375

Contents lists available at ScienceDirect

European Journal of Radiology

journal homepage: www.elsevier.com/locate/ejrad

Research article

Towards clinical implementation of an AI-algorithm for detection of cervical spine fractures on computed tomography

Huibert C. Ruitenbeek[a], Edwin H.G. Oei[a], Bart L. Schmahl[a], Eelke M. Bos[b], Rob J.C. G. Verdonschot[c], Jacob J. Visser[a,*]

[a] Department of Radiology and Nuclear Medicine, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, the Netherlands
[b] Department of Neurosurgery, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, the Netherlands
[c] Emergency Department, Erasmus MC, University Medical Center Rotterdam, P.O. Box 2040, 3000 CA Rotterdam, the Netherlands

ARTICLE INFO

Keywords:
Cervical spine
Artificial intelligence
Spinal fractures
Tomography spiral computed

ABSTRACT

Background: Artificial intelligence (AI) applications can facilitate detection of cervical spine fractures on CT and reduce time to diagnosis by prioritizing suspected cases.

Purpose: To assess the effect on time to diagnose cervical spine fractures on CT and diagnostic accuracy of a commercially available AI application.

Materials and methods: In this study (June 2020 - March 2022) with historic controls and prospective evaluation, we evaluated regulatory-cleared AI-software to prioritize cervical spine fractures on CT. All patients underwent non-contrast CT of the cervical spine. The time between CT acquisition and the moment the scan was first opened (DNT) was compared between the retrospective and prospective cohorts. The reference standard for determining diagnostic accuracy was the radiology report created in routine clinical workflow and adjusted by a senior radiologist. Discrepant cases were reviewed and clinical relevance of missed fractures was determined.

Results: 2973 (mean age, 55.4 ± 19.7 [standard deviation]; 1857 men) patients were analyzed by AI, including 2036 retrospective and 938 prospective cases. Overall prevalence of cervical spine fractures was 7.6 %. The DNT was 18 % (5 min) shorter in the prospective cohort. In scans positive for cervical spine fracture according to the reference standard, DNT was 46 % (16 min) shorter in the prospective cohort. Overall sensitivity of the AI application was 89.8 % (95 % CI: 84.2–94.0 %), specificity was 95.3 % (95 % CI: 94.2–96.2 %), and diagnostic accuracy was 94.8 % (95 % CI: 93.8–95.8 %). Negative predictive value was 99.1 % (95 % CI: 98.5–99.4 %) and positive predictive value was 63.0 % (95 % CI: 58.0–67.8 %). 22 fractures were missed by AI of which 5 required stabilizing therapy.

Conclusion: A time gain of 16 min to diagnosis for fractured cases was observed after introducing AI. Although AI-assisted workflow prioritization of cervical spine fractures on CT shows high diagnostic accuracy, clinically relevant cases were missed.

1. Introduction

Traumatic spinal fractures are severe injuries with a total global incidence of approximately 10.5 cases per 100.000 persons. In almost half of these cases the cervical spine is involved. Furthermore, the incidence of traumatic spine injury has gradually increased over the past two decades.[1–3] Early treatment of cervical spine injury leads to improved neurologic outcome compared to delayed treatment.[4] Therefore, timely diagnosis and treatment are crucial.

Computed tomography (CT) is the most suitable imaging modality to detect and characterize fractures of the cervical spine and, as a result, the application of CT in case of a suspected fracture is increasingly recommended in emergency medicine guidelines.[5] Consistent with the increasing incidence of traumatic spine injury, the amount of cervical spine CT scans also increases, which poses a burden on radiologists. In a trauma setting, cervical spine imaging is often part of a more

Abbreviations: AI, Artificial intelligence; NLP, Natural language processing; DNT, Detection and notification time; TAT, Turn-around-time.

* Corresponding author at: Department of Radiology and Nuclear Medicine, Erasmus MC, University Medical Center Rotterdam, Dr. Molewaterplein 40, Rotterdam, Zuid-Holland NL 3015 GD, the Netherlands.

E-mail address: j.j.visser@erasmusmc.nl (J.J. Visser).

https://doi.org/10.1016/j.ejrad.2024.111375
Received 29 September 2023; Received in revised form 9 February 2024; Accepted 15 February 2024
Available online 16 February 2024

extensive imaging protocol or even a full body CT scan. In this case, diagnosis of the most life-threatening conditions such as cervical spine fractures is important as this allows for rapid initiation of the most appropriate patient management.

Artificial intelligence (AI)-based radiology solutions are gaining ground for a wide range of clinical applications and various imaging modalities.[6] These applications are mostly focused at assisting or automating tasks to make optimal use of limited resources or personnel. Examples include task prioritization, image enhancement and computer-aided detection.[7] Thorough clinical validation and scientific evidence of their added value in clinical practice is needed before they can be clinically implemented on a large scale.[8].

Two articles on cervical spine fracture detection on CT using AI focused on diagnostic accuracy and found diverse results. [9–11] Missed cases often required stabilizing therapy during follow-up, demonstrating its limited clinical utility.[11] None of these studies reported on the effect on workflow and time to diagnosis.

Our objective was to assess the impact of a commercial AI algorithm on the time between scan acquisition and the radiologist assessment in cases involving cervical spine fractures on CT scans and we assessed diagnostic accuracy of the AI algorithm in a standalone setting (sensitivity and specificity) as secondary outcome.

2. Methods

This study was approved by the institutional review board of the Erasmus Medical Center, Rotterdam (Study ID: METC-2021-0685) and the obligation to obtain informed consent was waived according to the local policies. Prior to transmission to the AI algorithm, all information was anonymized, retaining solely the imaging data.

2.1. Study design and study population

A retrospective and a prospective cohort of patients were collected with the same inclusion criteria. 2500 patients were included consecutively in the retrospective cohort in the period from June 2020 until November 2021. The prospective cohort inclusion started in November 2021 and lasted until March 2022. Patients eligible for this study where 18 years and older who underwent any CT scan that included, but was not necessarily focused on, the cervical spine. The CT scans were acquired using a Siemens Multidetector CT (64-slice r higher), tubevoltage automatically selected and optimized based on body size and scan indication (70–140 kV), tubecurrent automatically adapted to patient size and selected kV, reconstruction matrix = 512*512 or higher, slice thickness/increment = 1.5 mm/1.5 mm, 3.0 mm/2.0 mm or 3.0 mm/3.0 mm, window width/level = 330/30, 350/50, 400/40, 600/100 or 3200/300. All CT scans containing the cervical spine used a protocol with axial acquisition and sagittal reformat. For some protocols intravenous contrast was administered. These scans were also included in the study cohort. Only studies containing multiplanar reformats in the sagittal plane were included as only these images could be processed by the algorithm. All scans were made at a tertiary care center, the Erasmus MC University Medical Center in Rotterdam, the Netherlands. We reported our results according to the Standards for the Reporting of Diagnostic accuracy studies (STARD) 2015 guidelines, which are recommended for reporting AI algorithms.[12].

2.2. AI algorithm

The AI product evaluated was Briefcase by AIDoc Medical (Briefcase version Rev. EU 9.0.0, AIDOC Medical, Tel Aviv, Israel). The workflow of this product entails the presentation of the entire CT scan, which needs to contain all cervical vertebrae to qualify for analysis, to the algorithm for further processing which does not require human input. All CT scans are screened by a quality check algorithm for field of view, reconstruction kernel and availability of multiplanar reconstructions in the sagittal plane with a maximum slice thickness of 5 mm (mm). The algorithm automatically selects the most suitable series, i.e. the sagittal reconstruction with the least amount of blur or artefacts, which is then analyzed for the presence or absence of fracture by the AI algorithm. The output of the algorithm is a binary value indicating either fractured or not fractured. Additionally, the algorithm provides the CT slice where the fracture was detected accompanied by a probability map of that slice.

All CT scans included in this study, both from the retrospective and prospective cohorts, were analyzed accordingly by the AI algorithm.

In addition, the AI product offers a support tool integrated in the radiology workflow, which is realized by installing a widget on the PACS workstations. Once installed, the product is constantly running in the background. If a fracture is detected a notification is pushed to all workstations to notify the reporting radiologists. This widget was not in effect for the retrospective cohort but enabled during the timeframe of the prospective cohort.

2.3. Detection and notification time

For every case in the study, both in the retrospective and prospective cohorts, a log file was retrieved from the PACS, containing timestamps of every moment the study was modified or viewed. For each case two timestamps were collected from these log files. The first timestamp was the moment the study was available in the PACS and for radiologists to review, the second timestamp was when the images were first opened by the radiologist. The amount of time between these two was defined as the ‘detection and notification time’ (DNT), a term that has been used in a recent study on the effect of AI on time to diagnosis in incidental pulmonary embolism detection.[13].

2.4. Reference standard

The reference standard was defined by the description of a fracture, or absence thereof, in the radiologist report. In the retrospective cohort a natural language processing (NLP) tool was used to extract information about the presence or absence of a fracture. The NLP tool produced ‘Positive’ as output when a fracture was described,’Negative’ if a fracture was ruled out or ‘Irrelevant’ if the report was inconclusive or incomplete. Radiology reports categorized as ‘Irrelevant’ were later assessed by an experienced musculoskeletal radiologist for categorization into ‘Positive’ or ‘Negative’ outcome. To assess the accuracy of the NLP algorithm, we retrieved a random sample as a method of quality assurance. Radiology reports of 50 cases where the NLP produced “Positive” as output and 50 cases where the NLP produced “Negative” as output were reviewed to find possible errors in the NLP result. In the prospective cohort the radiology report was assessed manually to assess whether a fracture was present or absent. The local incident registry was assessed to identify fractures initially missed by the attending radiologist that were detected later during treatment.

All radiologist reports were written or supervised by board-certified radiologists with subspecialty training in MSK. In case of discrepancy between the radiologist report and the algorithm result, the study was reviewed by an experienced musculoskeletal radiologist to define the reference standard. This review was performed on a clinical workstation with diagnostic screen and all relevant clinical information available. This included all original imaging series, medical history, clinical information and follow-up results.

2.5. Discrepant case analysis

Discrepancies between the radiology report and AI algorithm were assessed and categorized. All discrepant cases were assessed by a board-certified radiologist with 8 years of musculoskeletal subspecialty experience. For false negative cases we assessed if stabilizing therapy was performed by examining the patient record. For false positive cases the anatomic structure or abnormality marked as fracture by the algorithm was categorized.

2.6. *Statistical analysis*

To assess the performance of the algorithm we analysed the retrospective cohort. The diagnostic accuracy was determined by calculating the sensitivity, specificity, negative predictive value, positive predictive value and overall diagnostic accuracy[14], using Microsoft Excel (2016). Normality assessment through histograms, Q-Q plots, and the Kolmogorov-Smirnov test ($p < 0.05$) revealed a deviation from normal distribution, justifying the use of non-parametric statistics for subsequent analyses. We compared the median DNT (in minutes) of the retrospective cohort before introduction of AI to the DNT of the prospective cohort after introduction of AI, using a Mann-Whitney *U* test. SPSS statistical software was used (IBM Corp. IBM SPSS Statistics for Windows, Version 28.0 released 2021, IBM Corp, Armonk, NY, United States of America).[15]. P-values of < 0.05 were considered statistically significant.

3. Results

3.1. *Baseline characteristics*

A total of 3455 cervical spine CT scans were included in this study, of which 2500 in the retrospective cohort and 955 studies in the prospective cohort. 468 studies were excluded for various reasons. A flowchart according to STARD 2015 is shown in Fig. 1. 97 scans were excluded by the AI algorithm, no detailed information on this action could be retrieved from the manufacturer.

A total of 2974 patients were analyzed by the AI algorithm, of which 2036 studies in the retrospective cohort and 938 studies in the prospective cohort. Of all analyzed scans 2085 were dedicated cervical spine scans and 889 were total body scans. The total study population consisted of 137 clinical scans, 813 outpatient and 2024 emergency scans. No initially missed fractures were identified during assessment of the local incident registry. (Table 1).

3.2. *Detection and notification time*

In the total study population, DNT was 5 min shorter in the prospective cohort compared to the retrospective cohort ($p = 0.034$). In studies containing a fracture by reference standard, the median DNT of the prospective cohort was 16 min shorter compared to the retrospective cohort ($p = 0.01$). In studies not containing a fracture by reference standard the median DNT was 3 min shorter in the prospective cohort compared to the retrospective cohort, however this was not statistically significant ($p = 0.16$). (Table 2).

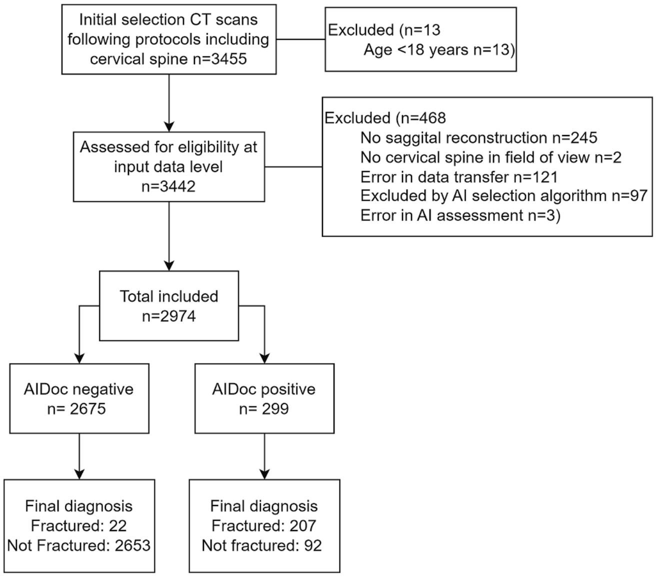
Initial selection CT scans following protocols including cervical spine n=3455
Excluded (n=13
Age <18 years n=13)
Assessed for eligibility at input data level n=3442
Excluded (n=468
No saggital reconstruction n=245
No cervical spine in field of view n=2
Error in data transfer n=121
Excluded by AI selection algorithm n=97
Error in AI assessment n=3)
Total included n=2974
AIDoc negative n= 2675
AIDoc positive n= 299
Final diagnosis Fractured: 22 Not Fractured: 2653
Final diagnosis Fractured: 207 Not fractured: 92

Fig. 1. Patient flow diagram conform STARD 2015 guidelines [12].

Table 1
Baseline characteristics of the study population.

	Mean age in years	Sex (%)	
		Male	Female
Total	55.4 ± 19.7 (range 18–101)	1857 (62.5 %)	1117 (37.5 %)
Prospective	56.8 ± 18.3 (range 18–95)	555 (59.2 %)	383 (40.8 %)
Retrospective	54.0 ± 20.0 (range 18–101)	1302 (64.0 %)	734 (36.0 %)

3.3. *Diagnostic accuracy*

In the retrospective cohort, 167 of 2036 (8.2 %) CT scans contained a fracture and 1869 of 2036 (91.8 %) scans did not contain a fracture defined by the radiology report. In the 100 radiology reports that were assessed to confirm the NLP output, no errors were observed. Out of 143 cases with discrepancies between AI algorithm and NLP, 108 were initially labelled positive and 35 were labelled negative for fracture by the by AI algorithm. The experienced musculoskeletal radiologist identified 42 as positive for fracture and 101 as negative. In one case the final diagnoses was changed from non-acute osteophyte fracture to injury related osteophyte fracture. In all other cases the conclusion of the AI algorithm was in accordance with the radiology report. Based on these findings the sensitivity of the algorithm was 89.8 % (95 % CI: 84.2–94.0 %) and the specificity was 95.3 % (95 % CI: 94.2–96.2 %). The resulting overall diagnostic accuracy was 94.8 % (95 % CI: 93.8–95.8 %). The negative predictive value (NPV) was 99.1 % (95 % CI: 98.5–99.4 %) and the positive predictive value (PPV) was 63.0 % (95 % CI: 58.0–67.8 %) (Table 3).

In the prospective cohort, 48 of 938 (5.1 %) CT scans contained a fracture and 890 of 938 (94.9 %) scans did not contain a fracture defined by the radiology report.

3.4. *Discrepant case analysis*

Fractures not detected by the algorithm were all located between vertebral levels C4 to C7. Missed fracture locations included the transverse process (9), articular process (3), spinous process (2), arch (1) and vertebral body (3). Of all missed fractures, five cases required stabilizing treatment during follow-up. These fractures were located at C4-5, C5, C6 and C7(2x). In two cases, missed fractures were invisible on the sagittal plane but only visible on the coronal plane due to an inadequate field of view and poor image quality (Table 4). An example shows sagittal and coronal reformats of the same patient. (Figs. 2 & 3).

In false positive cases (92 cases), the most common reasons were vascular canals (35 cases), ossifications or calcifications of the longitudinal ligaments (18 cases) and degenerative changes (including osteophytes, facet joint degeneration and ossification of the anterior and posterior longitudinal ligaments) (20 cases)(Table 4). These degenerative changes cause cortical discontinuity that was frequently marked as positive by the algorithm (Fig. 4). Some false positive cases were in fact based on fractures outside the cervical spine detected on CT scans that extended to other body parts. These were locations either in close proximity including one rib fracture and one skull base fracture, or in the pelvis (11 cases).

4. Discussion

In this study, we investigated the effect on detection and notification time of an AI-algorithm for the detection of cervical spine fractures on

Table 2
The median detection and notification time (DNT) was analyzed for fractured and not fractured cases. Times in the subgroup “Fractured” are from cases with fracture in the reference standard. Difference in medians was tested for significance using Mann-Whitney *U* test with $p < 0.05$.

	Median DNT (hours:minutes (IQR))						
	Total		Retrospective		Prospective		Significance
		#		#		#	
Overall	00:28 (00:11–01:49)	2344	00:28 (00:11–02:01)	1994	00:23 (00:08–01:18)	350	p = 0.034
Fractured	00:31 (00:13–04:53)	185	00:35 (00:13–06:15)	162	00:19 (00:02–00:28)	23	p = 0.01
Not fractured	00:28 (00:11–01:47)	2159	00:28 (00:11–01:53)	1832	00:25 (00:08–01:20)	327	p = 0.16

Table 3
Diagnostic accuracy: A total of 2036 retrospective studies were included. Total number of 238 positive cases were flagged by AIDoc.

	Fracture ground truth	No fracture ground truth	Total
Fracture AI algorithm	150	88	238
No fracture AI algorithm	17	1781	1798
Total	167	1869	2036

Table 4
False negative cases the follow-up information was assessed to find if stabilizing therapy was performed. For false positives the image aspect marked as fracture by the algorithm was categorized.

False negatives	Total	22 out of 2675 AI negative
	Fracture requiring stabilizing therapy	5
	Fracture not requiring stabilizing therapy	17
False positives	Total	92 out of 299 AI positive
	Vascular canal	35
	Degenerative	20
	Calcification/ossification	18
	Any finding outside cervical spine	11
	Old fracture/non-union	7
	Implants	1

CT in a patient cohort representative for clinical practice.

We found a time reduction for fractured cases in the prospective cohort of 16 min compared to the retrospective cohort as our primary outcome. This indicates that the implementation of this AI algorithm can be a successful triage tool and reduce the time to diagnosis of a cervical spine fracture. It is possible that this reduction in DNT is seen in the entire study population if, the general workload at the department was lower during the period after introduction of AI. To rule this out we have compared the DNT for fractured and non-fractured cases and found the difference not statistically significant. Therefore we conclude that a significant reduction in DNT of 16 min was seen in fractured cases after introduction of AI. Use of AI in the prospective cohort might consume extra time that has a risk of causing overestimation of the true DNT.

The detection and notification time was assessed instead of the turnaround-time(TAT) of the report. Following local hospital protocols, the radiologist contacted the treating physician immediately as soon as a cervical spine fracture was detected. This means that time of detection is well before the time that the report was finalized. Therefore, DNT is the most reliable parameter to assess the impact of the AI tool.[13] Analyzing the time of finalizing the report, or TAT, would underestimate the algorithm effect as reporting was often interrupted and reporting delayed.

Additionally, sensitivity was 89.8 % and specificity was 95.3 %. Three previous studies assessed the performance of the similar algorithm.[9–11] The study populations consisted of 1904, 665 and 2368 patients included consecutively. One study only included patients with an MRI within 48 h after CT acquisition.[10] The prevalence of fractures was 6.4 % [9], 21 % [10] and 9.3 % [11] compared to 7.6 % in our study. The resulting sensitivity of 54.9 % [9], 76 % [10] and 72 % [11] were lower compared to our results. A possible explanation of the higher sensitivity observed in our study could be that we used a newer version of the AI algorithm which was improved by additional model training. The most important difference between the three previous studies and ours is the definition of reference standard. In Small et al. and Voter et al., fracture presence was manually extracted from the report. In the study by Van den Wittenboer et al., the report also served as basis for the reference standard, but all images were re-read by a neuroradiologist. In all studies, including ours, discrepancies between the report and the algorithm were reviewed by readers not involved in the initial reporting. If the conclusion of the report was then adjusted, this final conclusion was used as reference standard. Our PPV (68.9 %) is within the range of the other studies (38.7 % [9], 87 % [10] and 85 % [11]). *Voter et al.* [9] analyzed discrepant cases and attributed false-positive results to degenerative damage and non-pathologic variants, comparable to our

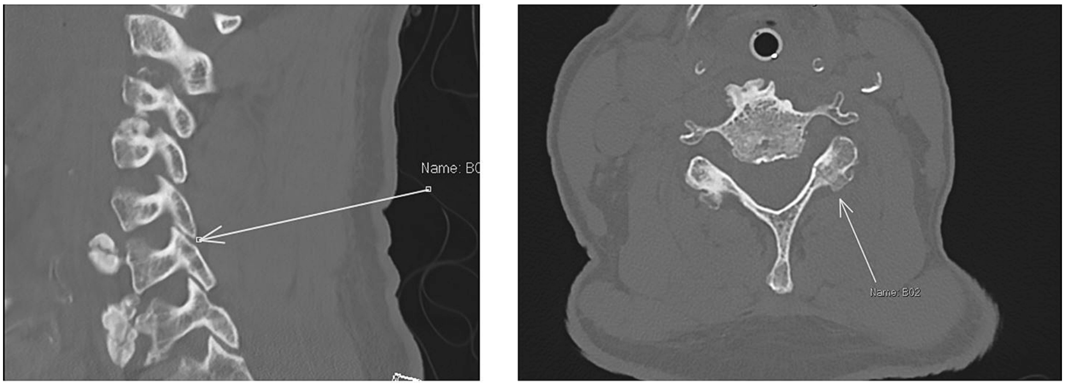

Fig. 2. Missed fractures by algorithm are sometimes poorly visible on the sagittal reformat while they are on the coronal reformat. Left: Sagittal reformat with a fractured transverse process. Right: Axial slice on which the fracture is visible.

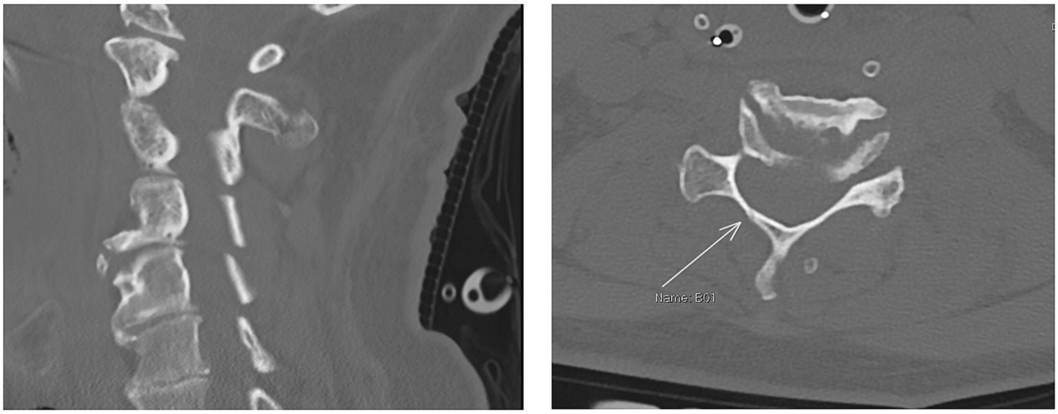

Fig. 3. Missed fracture of the vertebral lamina missed by the AI algorithm. On the sagittal reformat (left) this fracture is not visible, whereas it can be appreciated on the axial slice (right).

findings. The resulting specificity in our study (95.3 %) is comparable to the other three studies (94.1 % [9], 96 % [10] and 98.6 % [11]).

A high NPV is important from the perspective of the emergency physician. If a cervical spine fracture is ruled out, patients can be relieved from neck stabilizers which allows for treatment of other injuries It may also result in earlier transfer of the patient, thereby vacating the trauma room for new patients. From a neurosurgical perspective, focus lies on unstable fractures in need of stabilizing therapy. From this perspective, a missed stable fracture has no clinical consequence, but missed unstable fractures can lead to instability and damage to neural structures. Looking at the results in this light, future research should focus on enhancing the algorithm's ability to characterize fractures accurately, as this could greatly benefit both emergency and neurosurgical care. Given AI's potential for varied final use cases, exploring its capacity to distinguish between fracture types becomes even more critical, promising to refine diagnosis and treatment strategies across different clinical scenarios.

Radiologists should be educated about the limitations of the AI-applications before clinical implementation. We found that a large part of false positives was caused by normal anatomic structures, such as vascular channels and ligament calcifications. The algorithm's limitation to processing only sagittal reconstructions, leading to missed fractures that were identifiable in axial formats and the significant rejection of images without detailed analysis, hinders its clinical effectiveness and contributes to false negative cases. Furthermore, the AI product in our study utilizes a widget. The radiologist receives a notification immediately after the algorithm detects a fracture and is presented a probability map (Fig. 4). The probability map is an imprecise guide and still requires the radiologist to identify a fracture in the red area.

A number of limitations are present in this study. First, We were unable to collect complete timestamp data for all cases, potentially affecting result accuracy. However, as both cohorts were randomly selected using the same criteria, systematic bias is unlikely. Despite this, the missing data may have limited the precision of our findings. Second, we used data from a single tertiary care center. Third, we did not use a randomized study design. The number of non-ED scans in our study was small compared to ED scans. Therefore, our DNT comparison was more suitable to identify trends than to find statistically significant differences for non-ED patients. A non-ED patient cohort of similar size would allow for a better comparison of the DNT. Therefore, we suggest to perform a randomized controlled trial as follow-up research to obtain even higher levels of evidence in the use of AI applications. Fourth, we used an NLP to extract the radiologist conclusion from the text reports. In cases where both the NLP and the algorithm have made an error in analysing the image, additional FP and FN cases may be present. We did not find errors in the sample of radiology reports we reviewed as NLP quality assurance. Therefore, we assume the amount of NLP errors in the total set is limited. Also, the reference standard was based on a single human reader setting because this according to the standard clinical workflow. Although single reader settings do not provide a strong reference standard, we intentionally chose to compare the performance of AI to the clinical standard instead of aiming for a best possible reference standard. Fifth, our study was performed on data acquired from June 2020 until March 2022. During this period the healthcare system was confronted with the COVID-19 pandemic which had an influence on the number and type of patient admissions, staff presence and, therefore, workload. The number of patients admitted to the ED was lower which may have had an influence on the workload. However, because the individual urgent care setting remains the same we expect that the result for the ED cohort would not be different. For non-ED cases it can be imagined that during this period workload of the radiology department was not comparable to regular clinical practice.

As, to our knowledge, this study is only the fourth evaluation of this specific AI tool, more validation studies are needed. These should report on diagnostic accuracy, time to diagnosis, and reporting times, but also take into account the impact on diagnostic thinking and therapeutic decision-making.[16] Ultimately, it should be clear what impact this tool has on clinically relevant parameters, such as length of stay in the hospital. In addition, implementation should be dependent on the cost-effectiveness of the tool, requiring specific research in this area. Trials conducted in a clinically representative population to demonstrate how these results translate into clinical practice are scarce.[17].

In conclusion, our study shows that the AIDoc algorithm significantly reduces turnaround time in emergency department settings, highlighting its efficiency in clinical workflow. While the AI's diagnostic performance shows limited positive predictive value, its high specificity and negative predictive value suggest a strong capability in ruling out fractures. Despite the need for further research for comprehensive evaluation, these findings indicate the potential of the AIDoc algorithm as an effective triage tool to expedite diagnosis in suspected cervical spine fracture cases.

Clinical relevance statement

Although AI applications have the potential to be used to reduce report turnaround time and facilitate early detection by prioritizing cases, our study showed high false negative numbers. Therefore, current evidence shows limited applicability for clinical practice.

CRediT authorship contribution statement

Huibert C. Ruitenbeek: Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Conceptualization. **Edwin H.G. Oei:** Writing – review & editing,

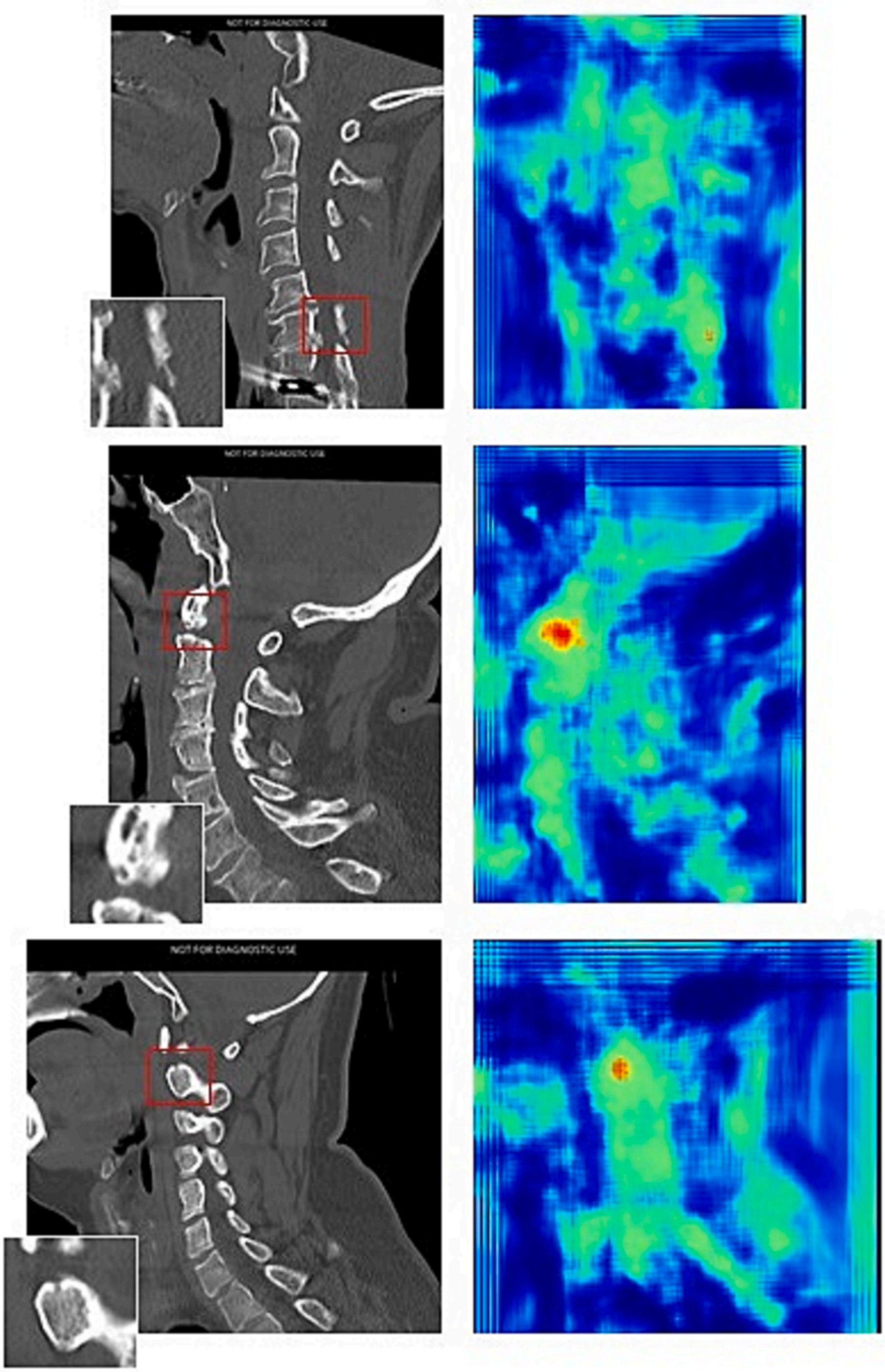

Fig. 4. Three images of false positive AI results. Images on the left show slices where a fracture has been detected, indicated by the red circle. Images on the right show a probability map of the same slice, a red color indicating presence of a fracture. Top row: Ligamentous calcification. Middle row: Degenerative change. Bottom row: Vascular channel.

Supervision, Methodology, Formal analysis, Conceptualization. **Bart L. Schmahl:** Writing – original draft, Project administration, Methodology, Conceptualization. **Eelke M. Bos:** Writing – review & editing, Conceptualization, Investigation, Validation, Methodology. **Rob J.C.G. Verdonschot:** Writing – review & editing, Conceptualization, Investigation, Validation, Methodology. **Jacob J. Visser:** Writing – review & editing, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Authors HR, BS, EMB, RV and EO hereby declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Author Jacob J. Visser conflicts of interest are the following: Medical advisor AstraZeneca (PINPOINT), Medical advisor NLC Ventures, Medical advisor Noaber Ventures, Medical advisor Contextflow, Medical advisor

Quibim, Medical advisor Tegus..

References

[1] P.G. Passias, et al., Traumatic Fractures of the Cervical Spine: Analysis of Changes in Incidence, Cause, Concurrent Injuries, and Complications Among 488,262 Patients from 2005 to 2013, World Neurosurg. 110 (2018) E427–E437.

[2] R. Kumar, et al., Traumatic Spinal Injury: Global Epidemiology and Worldwide Volume, World Neurosurg. 113 (2018) e345–e363.

[3] N.C. Utheim, et al., Epidemiology of traumatic cervical spinal fractures in a general Norwegian population, Injury. Epidemiol. 9 (1) (2022).

[4] M.G. Fehlings, et al., Early versus Delayed Decompression for Traumatic Cervical Spinal Cord Injury: Results of the Surgical Timing in Acute Spinal Cord Injury Study (STASCIS), PLoS One 7 (2) (2012).

[5] E. Gesu, et al., Management of patients with cervical spine trauma in the emergency department: a systematic critical appraisal of guidelines with a view to developing standardized strategies for clinical practice, Intern. Emerg. Med. 16 (8) (2021) 2277–2296.

[6] K.G. van Leeuwen, et al., Artificial intelligence in radiology: 100 commercially available products and their scientific evidence, Eur. Radiol 31 (6) (2021) 3797–3804.

[7] K.G. van Leeuwen, et al., How does artificial intelligence in radiology improve efficiency and health outcomes? Pediatr. Radiol (2021).

[8] F.L. Jacobson, E.A. Krupinski, Clinical Validation Is the Key to Adopting AI in Clinical Practice, Radiol. Artif. Intell 3 (4) (2021) e210104.

[9] A.F. Voter, et al., Diagnostic accuracy and failure mode analysis of a deep learning algorithm for the detection of cervical spine fractures, AJNR Am. J. Neuroradiol 42 (8) (2021) 1550–1556.

[10] J.E. Small, et al., CT cervical spine fracture detection using a convolutional neural network, AJNR Am. J. Neuroradiol 42 (7) (2021) 1341–1347.

[11] G.J. van den Wittenboer, et al., Diagnostic accuracy of an artificial intelligence algorithm versus radiologists for fracture detection on cervical spine CT, Eur. Radiol (2024).

[12] P.M. Bossuyt, et al., STARD 2015: An Updated list of essential items for reporting diagnostic accuracy studies, Radiology 277 (3) (2015) 826–832.

[13] L. Topff, et al., Artificial intelligence tool for detection and worklist prioritization reduces time to diagnosis of incidental pulmonary embolism at CT, Radiol.: Cardiothor. Imaging 5 (2) (2023) e220163.

[14] G. Grunau, S. Linn, Detection and diagnostic overall accuracy measures of medical tests, Rambam Maimonides Med. J. 9 (4) (2018).

[15] IBM Corp., *IBM* SPSS Statistics for Windows, Version 28.0.1.0. 2021.

[16] D.G. Fryback, J.R. Thornbury, The efficacy of diagnostic-imaging, Med. Decis. Making 11 (2) (1991) 88–94.

[17] M. Nagendran, et al., Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies, Bmj-British. Med. J. (2020) 368.