Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Cervical Spine Fractures

A.F. Voter, M.E. Larson, J.W. Garrett, and J.-P.J. Yu

ABSTRACT

BACKGROUND AND PURPOSE: Artificial intelligence decision support systems are a rapidly growing class of tools to help manage ever-increasing imaging volumes. The aim of this study was to evaluate the performance of an artificial intelligence decision support system, Aidoc, for the detection of cervical spinal fractures on noncontrast cervical spine CT scans and to conduct a failure mode analysis to identify areas of poor performance.

MATERIALS AND METHODS: This retrospective study included 1904 emergent noncontrast cervical spine CT scans of adult patients (60 [SD, 22] years, 50.3% men). The presence of cervical spinal fracture was determined by Aidoc and an attending neuroradiologist; discrepancies were independently adjudicated. Algorithm performance was assessed by calculation of the diagnostic accuracy, and a failure mode analysis was performed.

RESULTS: Aidoc and the neuroradiologist's interpretation were concordant in 91.5% of cases. Aidoc correctly identified 67 of 122 fractures (54.9%) with 106 false-positive flagged studies. Diagnostic performance was calculated as the following: sensitivity, 54.9% (95% CI, 45.7%–63.9%); specificity, 94.1% (95% CI, 92.9%–95.1%); positive predictive value, 38.7% (95% CI, 33.1%–44.7%); and negative predictive value, 96.8% (95% CI, 96.2%–97.4%). Worsened performance was observed in the detection of chronic fractures; differences in diagnostic performance were not altered by study indication or patient characteristics.

CONCLUSIONS: We observed poor diagnostic accuracy of an artificial intelligence decision support system for the detection of cervical spine fractures. Many similar algorithms have also received little or no external validation, and this study raises concerns about their generalizability, utility, and rapid pace of deployment. Further rigorous evaluations are needed to understand the weaknesses of these tools before widespread implementation.

ABBREVIATIONS: AI = artificial intelligence; ASiR = adaptive statistical iterative reconstruction; DSS = decision support system; CSFx = cervical spinal fractures

Cervical spinal fractures (CSFx) are devastating injuries that can cause severe morbidity and mortality from damage to the enclosed spinal cord, the craniocervical junction, and cervical vasculature.[1] Failure of the osseous spinal column can lead to instability and impingement of the underlying spinal cord;[2] therefore, timely identification and stabilization of CSFx are crucial to prevent further disability.[1,3] In the acute clinical setting, NCCT of the cervical spine is the recommended method for detecting CSFx;[4] however, with diagnostic imaging volumes dramatically increasing,[5,6] these increased imaging volumes place a burden on radiologists who must maintain diagnostic accuracy and efficiency.[7] While there has been great effort to reduce the number of unnecessary scans ordered, including the use and implementation of the National Emergency X-Radiography Utilization Study Group[8] criteria and the Canadian C-Spine Rule[9] to reduce the number of unnecessary cervical spinal NCCTs, their effectiveness appears to be modest,[10,11] and diagnostic imaging volumes continue to increase.

To assist radiologists in managing these rising case volumes, artificial intelligence (AI) decision support systems (DSSs) have been developed to help prioritize imaging studies with critical

Received December 9, 2020; accepted after revision March 14, 2021.

From the School of Medicine and Public Health (A.F.V.) and Departments of Radiology (M.E.L., J.W.G., J.-P.J.Y.), Biomedical Engineering (J.-P.J.Y.), College of Engineering, and Psychiatry (J.-P.J.Y.), University of Wisconsin School of Medicine and Public Health, Madison, Wisconsin.

A.F.V. was supported by National Institutes of Health F30 CA210465 and T32 GM008692; J.W.G. was supported by National Institutes of Health R01 LM013151; and J.-P.J.Y. was supported by the Clinical and Translational Science Award program, through the National Institutes of Health National Center for Advancing Translational Sciences, grant UL1TR002373.

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Please address correspondence to John-Paul J. Yu, MD, PhD, Department of Radiology, University of Wisconsin-Madison, 600 Highland Ave, D4-352, Madison, Wisconsin; e-mail: jpyu@uwhealth.org

Indicates open access to non-subscribers at www.ajnr.org

Indicates article with online supplemental data.

http://dx.doi.org/10.3174/ajnr.A7179

findings.[12,13] These DSSs identify and subsequently flag studies with actionable results, allowing radiologists to prioritize them over scans with likely negative findings to speed the reporting of critical findings. However, DSSs that incorrectly flag an excessive number of studies with negative findings or conversely miss critical findings might slow the radiologist's performance. Rigorous analysis is, therefore, crucial. AI algorithms are known to have numerous limitations, including the need for large, diverse, and unbiased datasets,[14] which can be difficult to acquire or curate[15] and operate in a manner that precludes direct interrogation of the decision process itself. These issues can lead to poor performance, which is difficult or impossible to troubleshoot, especially when the algorithms are implemented in settings beyond their initial training environment.[16–18] While the rapid development and clinical implementation of DSSs are exciting, this proliferation risks outstripping our ability to rigorously assess and validate their performance. This validation and assessment have not been extensively performed or reported in the literature. Furthermore, site-specific performance differences without obvious etiologies have been observed for AI DSSs.[16-18] Thus, rigorous studies to guide AI DSS installations in varied clinical settings and a greater understanding of the generalizability (or lack thereof) of AI DSSs are needed to safely translate this important tool into widespread clinical practice.

Our institution recently implemented Aidoc (Aidoc Medical), an FDA-cleared, commercially available AI DSS for the detection of CSFx.[19] While several spine fracture DSSs have been developed,[19-23] their diagnostic accuracy and overall performance remain unknown. To gain insight into the performance of this system specifically and AI DSSs more generally, we conducted a retrospective review of Aidoc as clinically implemented in our institution. The aim of this study was to characterize the performance of Aidoc for the detection of CSFx and conduct a failure mode analysis to identify areas of poor diagnostic performance.

MATERIALS AND METHODS

This Health Insurance Portability and Accountability Act–compliant retrospective study was approved by the institutional review board. The requirement for informed consent was waived. The data were analyzed and controlled by the authors exclusively, none of whom are employees of or consultants to Aidoc Medical or its competitors.

Study Population, Data Collection, Imaging Parameters, and AI System

Adult (older than 18 years of age) CT cervical spine studies without contrast from January 20, 2020, to October 8, 2020, in our radiology information system were identified and contemporaneously processed by Aidoc. Pediatric (younger than 18 years of age) studies and examinations with intrathecal contrast were excluded from this study. Scans were performed at an academic level I trauma center and associated outreach imaging centers with a fleet of 9 models of scanners (GE Healthcare) (summarized in Online Supplemental Data). A total of 1904 adult, noncontrast cervical spine CT scans were identified in 1923 emergent neck CT scans (mean age, 60 [SD, 22] years; 50.3% men). Acquisition parameters for noncontrast CT examinations of the cervical spine are as follows: 120 kV(peak); axial helical acquisition; pitch = 0.625 mm; rotation speed = 5.6 mm/rotation; rotation time = 0.5 seconds; automatic exposure control = smart mA (230–750 mA); section thickness = 1.25 mm; interval = 0.625 mm. Standard soft-tissue and bone window (Bone Plus algorithm [GE Healthcare]) reconstructions were contemporaneously generated for review by radiologists (1.25-mm section thickness, sagittal and coronal; 0.625-mm interval; no adaptive statistical iterative reconstruction [ASiR]). Immediately following study acquisition, axial thin bone (Bone Plus reconstruction; 0.625-mm section thickness; 0.312-mm interval; no ASiR) and sagittal bone (Bone Plus reconstruction; 1.5-mm section thickness; 0.98-mm interval; no ASiR) series were generated and analyzed by the Aidoc algorithm, which then classifies each scan as positive or negative for CSFx. Aidoc-specific image series were not available to the interpreting radiologist for review. However, because the algorithm was evaluated as clinically implemented, the final Aidoc classification and key image indicating the flagged pathology were available to the radiologist at the time of initial study interpretation. For the purposes of this study, the final neuroradiologist interpretation serves as ground truth data and is in keeping with prior approaches evaluating the diagnostic performance of AI-related systems.[24,25]

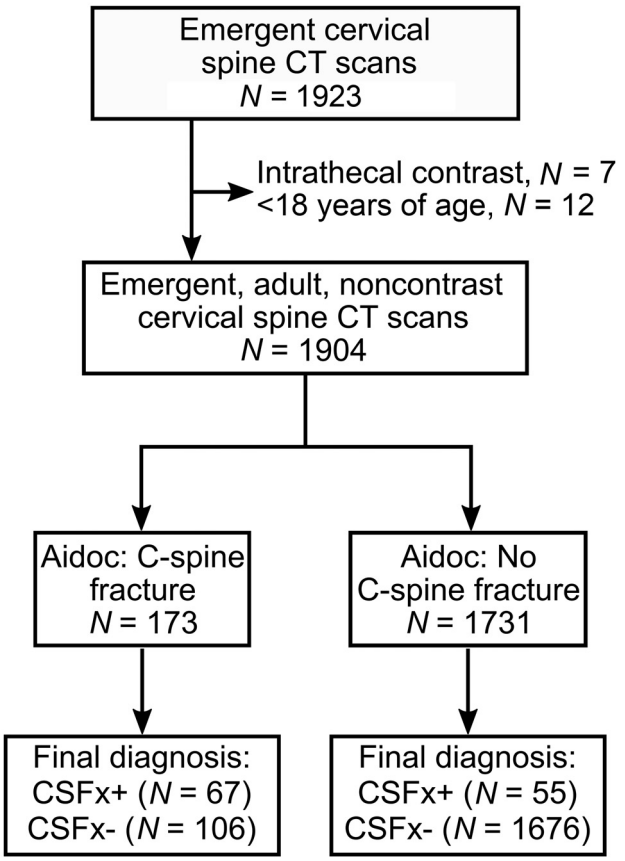

FIG 1. Standards for reporting diagnostic accuracy studies (STARD) patient flow diagram.

Data Processing and Analysis

The presence of a cervical spine fracture, type of fracture, vertebra fractured, estimate of fracture age, and study indication were manually extracted from the attending neuroradiologist imaging

Table 1: The impact of patient characteristics on Aidoc performance

Factor		Aidoc Incorrect (No.)	%	*P* Value
Total (No.) (%)	1904, 100	161	100	
Indication (No.) (%)				.97
Trauma	1796, 94	155	96	
Critical	511, 27	45	28	
Minor	888, 47	74	46	
Not specified	397, 21	36	22	
Neck pain	27, 1	1	1	
Neurologic deficit	33, 2	2	1	
Postoperative	10, 1	1	1	
Other	38, 2	2	1	
Sex (No.) (%)				.08
Male	958, 50	92	57	
Female	946, 50	69	43	
Imaging location (No.) (%)				.86
Academic center	1659. 87	141	88	
Outreach center	245, 13	20	12	
History of cervical spine surgery (No.) (%)				.57
Prior surgery	67, 4	7	4	
No prior surgery	1837, 96	154	96	
Age (mean) (yr)				.03
Overall	60 (SD, 22)			
Aidoc incorrect	64 (SD, 21)			
Aidoc correct	60 (SD, 22)			

Table 2: Etiology of the false-positive flagged studies

False-Positive Etiology	Count	Percentage of All Flagged Studies ($n = 173$)
Degeneration	55	31.8
Degenerative ossicle	18	10.4
Facet degeneration	14	8.1
Calcified ligament	6	3.5
Cortical irregularity	7	4.0
Osteopenia	4	2.3
Cystic degeneration	4	2.3
Atlantodental joint	1	0.6
Osteophyte	1	0.6
Noncervical pathology	15	8.7
Rib fracture	8	4.6
Degeneration, thoracic	4	2.3
Skull fracture	2	1.2
Carotid calcification	1	0.6
Anatomic variant	10	5.8
Nonunion vertebrae	4	2.3
Transitional anatomy	2	1.2
Limbus	2	1.2
Bifid spinous process	1	0.6
Secondary transverse foramen	1	0.6
Nutrient foramen	9	5.2
Artifact	7	4.0
Unknown	8	4.6
Other	2	1.7
DISH	1	0.6
Occipital suture	1	0.6
Total	106	61.3

Note:—DISH indicates diffuse idiopathic skeletal hyperostosis.

report of each study. To establish the ground truth of the presence or absence of an CSFx, we compared the interpretations of the neuroradiologist and Aidoc. Concordant interpretations were assumed to be correct; studies with discordant interpretations were reviewed by a third independent reviewer not involved in the initial interpretation (radiology resident and attending neuroradiologist with 6 years of experience) to make a final ground truth determination. Study indication was inferred from the report body and imaging order. Critical traumas included motor vehicle collisions, falls from heights or stairs, sporting accidents, assaults, and hangings. Minor traumas largely involved falls from standing height or lower. Last, traumas were categorized as "not specified" if there was insufficient information regarding the mechanism of trauma.

Statistical Analysis

χ^2 tests and 2-sided paired *t* tests were used for statistical testing for categoric and quantitative comparisons, respectively, with a significance threshold of .05. Diagnostic accuracy (sensitivity, specificity, positive predictive value, negative predictive value, and tests for statistical significance were all performed in Excel 365 [Microsoft]).

RESULTS

To gauge the diagnostic accuracy of Aidoc as clinically implemented in our institution, we identified 1904 noncontrast cervical spine CTs for inclusion during our study. A total of 173 (9.1%) of the total studies were flagged by Aidoc as positive for CSFx, and CSFx were identified on 38.7% (67/173) of the flagged studies. Of the studies not flagged by Aidoc, 3.2% (55/1731) contained fractures (Fig 1). Diagnostic performance characteristics with 95% confidence intervals were determined as follows: sensitivity, 54.9% (95% CI, 45.7%–63.9%); specificity, 94.1% (95% CI, 92.9%–95.1%); positive predictive value, 38.7% (95% CI, 33.1%–44.7%); and negative predictive value, 96.8% (95% CI, 96.2%–97.4%).

First, we sought to understand how patient factors might impact the diagnostic accuracy of Aidoc (Table 1). Because the mechanism of injury can determine the type and severity of injury, we calculated the Aidoc false-negative rate based on the indication for the CT examination of the cervical spine (eg, trauma, neck pain, neurologic deficit). No significant differences in Aidoc performance were noted for any of the study indications, study location (ie, academic center or outreach imaging center), or model of CT scanner (Online Supplemental Data, $P = .82$). Similarly, the diagnostic error rate of Aidoc was not impacted by either patient sex or history of cervical spine surgery. We did observe, however, that patients incorrectly classified by Aidoc were older than those correctly classified (mean 64 [SD, 21] years versus 60 [SD, 22] years, respectively; $P = .03$).

Next, we examined whether characteristics of the individual fractures impacted algorithm performance (Online Supplemental Data). Aidoc performance was found to be independent of the number of vertebrae fractured (single versus multiple) and the identity of the fractured vertebrae. However, while they were not significant as a category, we observed a lower rate of incorrect Aidoc calls with injuries of C2 and a higher rate at C5. We also

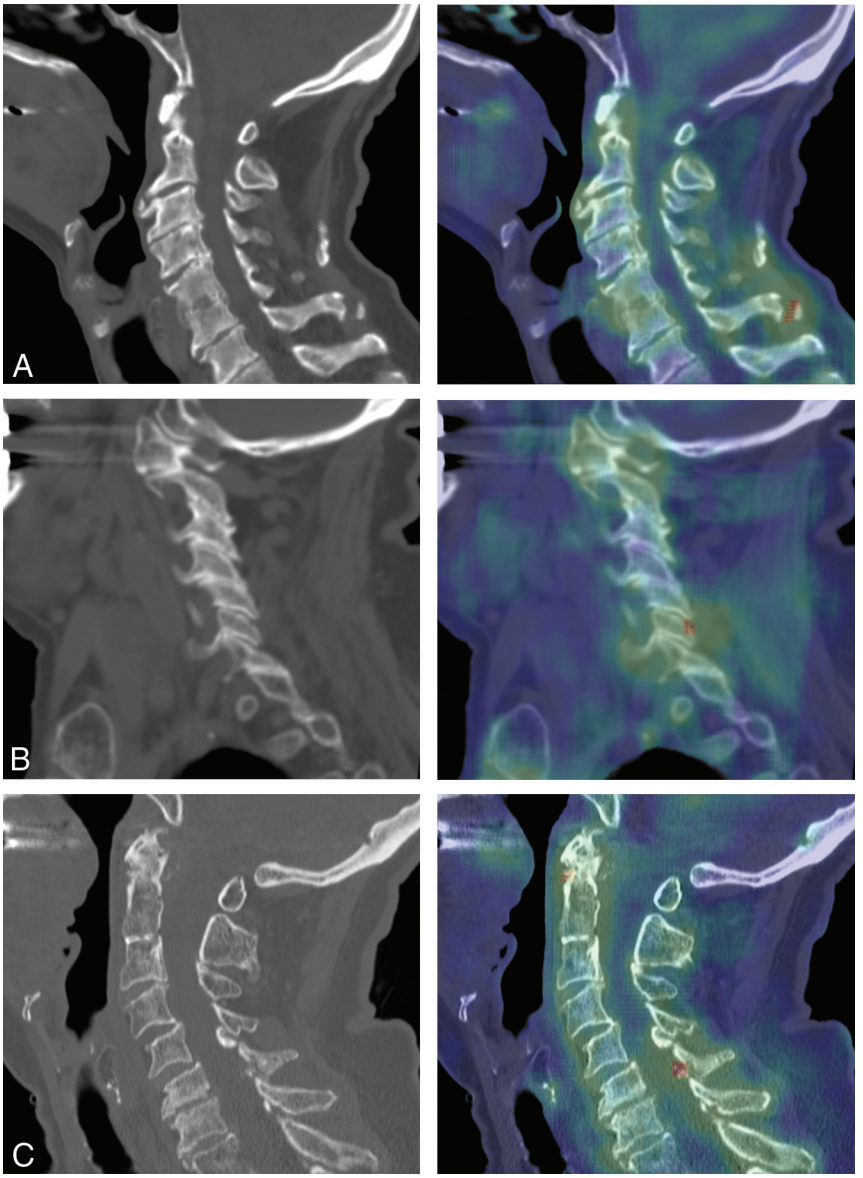

FIG 2. Examples of degenerative findings falsely flagged by Aidoc. Each panel shows the sagittal noncontrast cervical spine CT (*left*) and the Aidoc key image indicating the flagged pathology in red (*right*). *A*, A chronic ossicle falsely flagged by Aidoc. *B*, False-positive findings triggered by facet degeneration. *C*, Ossification of the ligamentum flavum incorrectly identified as a fracture by Aidoc.

observed that the algorithm was significantly more successful at identifying acute fractures than nonacute fractures (ie, chronic or age-indeterminate). Furthermore, location of the fracture within each vertebra was a significant contribution to algorithm performance, with fractures of osteophytes or the vertebral body overrepresented in the false-negative studies.

The timely identification of new fractures is of particular clinical importance, so we explored the performance of Aidoc in the detection of acute fractures. We did not find any significant differences between the acute fractures correctly flagged by Aidoc and those it missed, though our analysis was limited by the relatively small number of acute fractures (Online Supplemental Data). However, the algorithm missed 50% (5 of 10) of acute fractures involving the transverse foramen.

Because the number of false-positive flagged studies exceeded the number of true-positives (106 versus 67), we next sought to understand the poor positive predictive value of Aidoc by exploring possible failure modes of the false-positive studies. Each study flagged by Aidoc is accompanied by a probability heat map highlighting the suspected fracture identified by Aidoc, thus allowing us to identify the etiology of each false-positive finding (Table 2). The most common etiology was the presence of degenerative structures such as a degenerative ossicle (Fig 2*A*), facet degeneration (Fig 2*B*), ossification of the ligamentum flavum (Fig 2*C*), or other degenerative cortical irregularities. The next most common sources of false-positive findings were pathologies outside the cervical spine and scope of the algorithm, such as rib or skull fractures (Fig 3*A*), and nonpathologic anatomic variants (Fig 2*B*, -*C*). False-positives were also found to have been triggered by motion artifacts or normal anatomy, and in a small number of cases, we were unable to identify any abnormality.

DISCUSSION

A wide range of AI DSSs have been developed to reduce the risk of missing or delaying the reporting of time-sensitive findings.[12,13] However, AI algorithms are known to have limitations and can be difficult to generalize to clinical sites with disease prevalence and imaging protocols that differ from training datasets. Because poorly performing DSSs can hinder radiologists, it is crucial that these tools undergo rigorous evaluation before widespread implementation. While the implementation of Aidoc for CSFx has excellent reported diagnostic characteristics (sensitivity of 91.7% and specificity of 88.6%, as reported in the initial FDA disclosure),[19] to our knowledge, no independent evaluations of its performance have been published or, more generally, any data evaluating the diagnostic accuracy of AI DSSs in detecting cervical spine fractures. To this end, we conducted a retrospective study to evaluate the diagnostic accuracy of Aidoc, an FDA-cleared AI DSS for the evaluation of CSFx as clinically implemented at our institution.

At our institution, Aidoc fared poorly, with a notably lower sensitivity and positive predictive value than initially reported to the FDA.[19] To understand this unexpected

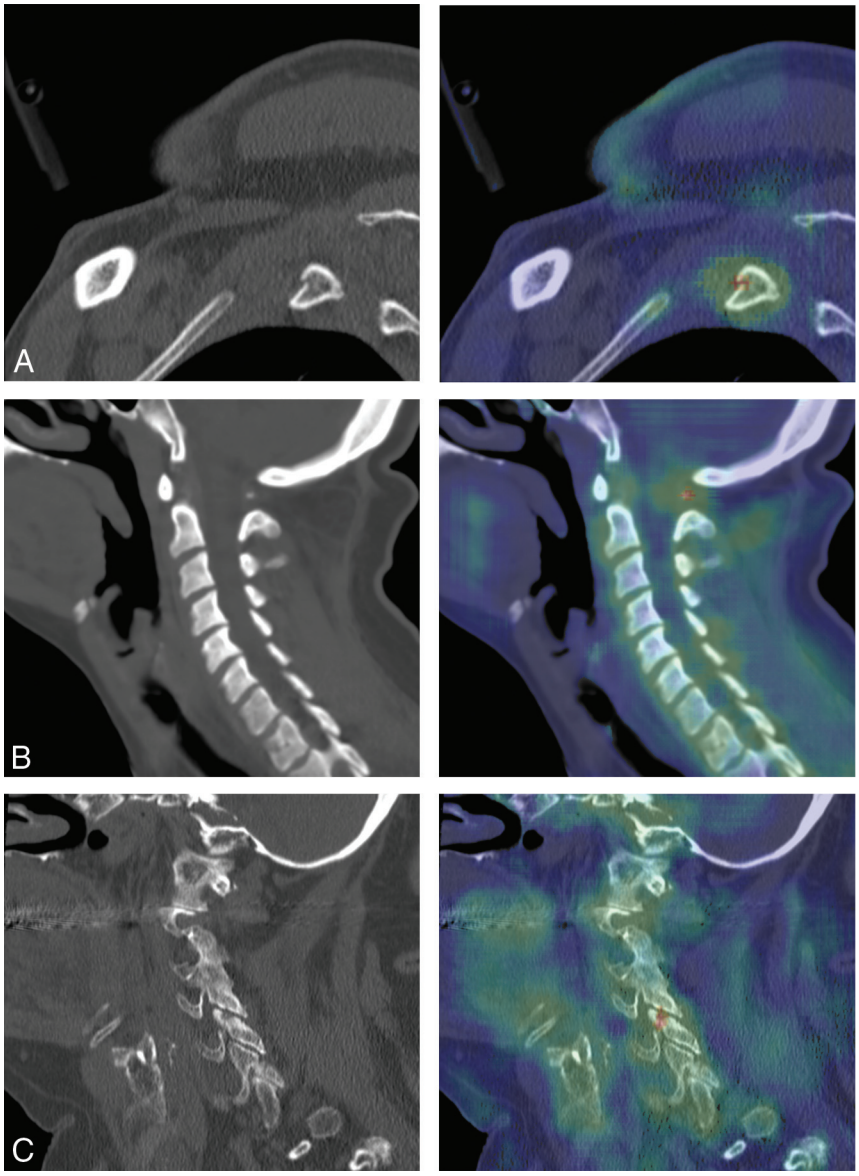

FIG 3. Examples of nondegenerative findings falsely flagged by Aidoc. Sagittal noncontrast cervical spine CT (*left*) and the Aidoc key image indicating the flagged pathology in red (*right*). *A*, Rib fracture outside of the cervical spine incorrectly flagged by Aidoc. *B*, Congenital hypoplasia of the posterior arch of the atlas flagged as a fracture. *C*, A nonpathologic nutrient foramen with degenerative changes identified as a fracture by Aidoc.

performance gap, we conducted a failure mode analysis to identify possible sources of this impaired performance. Neither imaging location, scanner model, nor study indications were found to be significantly associated with the diagnostic performance of Aidoc. However, the sensitivity was affected by patient age and characteristics of the underlying fracture, specifically the fracture acuity and location of the fracture, with chronic fractures and fractures of osteophytes and the vertebral body overrepresented among the missed fractures. Osteophyte formation and compression fractures are degenerative in nature, so underperformance in their detection may contribute to the worsened algorithm performance in older patients.

Because the value of this and similar algorithms stems from the faster detection of findings that can alter clinical management, it is especially important to consider the performance in the detection of acute fractures. We did not find any differences between the acute fractures correctly identified or missed by Aidoc, though our statistical analysis was limited by the relatively small number of acute fractures missed by the algorithm. However, it is notable that the 50% of the acute fractures involving the transverse foramen were missed by Aidoc. These fractures can indicate compromise of the underlying vertebral artery, so rapid detection by the algorithm is especially valuable and more examples should be included in the algorithm training set.

In cases with multiple fractures, the algorithm needs to correctly identify only a single fracture to score as correct. Therefore, we hypothesized that these studies would have a lower false-negative rate. However, we observed that the miss rate did not depend on the total number of fractures present in an imaging examination, suggesting that fracture identification may have been precluded by other features of the study rather than fracture characteristics themselves.

We noted a significant and unexpected number of false-positive studies in our dataset, outnumbering the flagged true CSFx. Spine degeneration was the most common etiology of false-positives observed. This is perhaps not surprising because degeneration occurs with aging and generates abnormalities such ossicles or irregularities in the bony surface that could be mistaken for fractures. Accordingly, the age of patients misclassified by Aidoc was higher than that in the correctly classified group, and we hypothesize that the increased burden of degeneration may have led to impaired performance. Our dataset lacked an accessible way to assess the extent of degeneration directly, but this could be explored in future studies. We speculate that greater representation of nonfractured examples of both degeneration and anatomic variants in the training set would likely reduce the false-positive burden, given their overrepresentation here in our analysis as false-positives. In addition, differences in diagnostic accuracy may also be attributed to institution-specific differences and would be difficult to disentangle. However, in the FDA 510(k) application, the number of cases positive and negative for CSFx were adjusted to be roughly equal. Because

diagnostic performance is strongly influenced by disease prevalence, this also likely contributes to the observed differences in the reported diagnostic accuracy of Aidoc and our clinical observations.[19] Our observed rate of positive findings is 6.4%, which reflects the true rate of CSFx at our institution. Because positive and negative predictive values depend on the underlying prevalence of the disease, we believe our measurements will more closely reflect the experience of other users. This discrepancy highlights an emerging need to standardize study design to allow rigorous and unbiased comparisons across different sites and for accurate reporting and evaluation of AI DSS algorithms in the imaging literature.

Our study has limitations that must be considered. First, because Aidoc has already been clinically implemented at our institution, the interpretation by Aidoc of each study was available to the neuroradiologist during the initial read. While this may have inflated the accuracy of the neuroradiologist's read, the diagnostic accuracy of Aidoc is unaffected. Additionally, while the Aidoc algorithm is available to all radiologists at our institution, there is marked variation in how it has been incorporated into their individual workflow. We were, therefore, unable to assess whether the algorithm reduced time to image analysis in cases flagged for CSFx. Nevertheless, given the poor positive predictive value, we suspect that any time savings would be diluted by the number of false-positives. Last, this single-institution study was performed at an academic center equipped with GE Healthcare scanners, potentially limiting the generalizability of our findings to institutions in other practice settings or those with a different fleet of scanners from other vendors.

CONCLUSIONS

We examined the diagnostic performance of Aidoc for the detection of CSFx as implemented at our institution and observed meaningful worse diagnostic accuracy than previously reported. Although the nature of neural network algorithms obscures a full understanding of this impairment, our failure mode analysis has identified several potential areas for improvement. Nevertheless, the overall performance of this AI DSS at our institution is different enough and raises potential concerns about the generalizability of AI DSSs across heterogeneous clinical environments and motivates the creation of data-reporting standards and standardized study design, the lack of which precludes unbiased comparisons of AI DSS performance across both institutions and algorithms. Adoption of a standardized design for all AI DSS algorithms will help speed the development and safe implementation of this promising technology as we aim to integrate this important tool into clinical workflow.

Disclosures: Andrew F. Voter—*RELATED: Grant*: National Institutes of Health.* John W. Garrett—*RELATED: Grant*: National Institutes of Health.* John-Paul J. Yu—*RELATED: Grant*: National Institutes of Health.* *Money paid to the institution.

REFERENCES

1. Copley D, Tilliridou D, Jamjoom M. **Traumatic cervical spine fractures in the adult.** *Br J Hosp Med (Lond)* 2016;77:530–35 CrossRef Medline
2. Denis F. **The three-column spine and its significance in the classification of acute thoracolumbar spinal injuries.** *Spine (Phila Pa 1976)* 1983;8:817–31 CrossRef Medline
3. Fischer PE, Perina DG, Delbridge TR, et al. **Spinal motion restriction in the trauma patient: a joint position statement.** *Prehosp Emerg Care* 2018;22:659–61 CrossRef Medline
4. Beckmann NM, West OC, Nunez D Jr, et al. Expert Panel on Neurological Imaging and Musculoskeletal Imaging. **ACR Appropriateness Criteria® Suspected Spine Trauma.** *J Am Coll Radiol* 2019;16:S264–85 CrossRef Medline
5. Hess EP, Haas LR, Shah ND, et al. **Trends in computed tomography utilization rates: a longitudinal practice-based study.** *J Patient Saf* 2014;10:52–58 CrossRef Medline
6. Kocher KE, Meurer WJ, Fazel R, et al. **National trends in use of computed tomography in the emergency department.** *Ann Emerg Med* 2011;58:452–62 CrossRef Medline
7. McDonald RJ, Schwartz KM, Eckel LJ, et al. **The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload.** *Acad Radiol* 2015;22:1191–98 CrossRef Medline
8. Hoffman JR, Mower WR, Wolfson AB, et al. **Validity of a set of clinical criteria to rule out injury to the cervical spine in patients with blunt trauma: National Emergency X-Radiography Utilization Study Group.** *N Engl J Med* 2000;343:94–99 CrossRef Medline
9. Mower WR, Wolfson AB, Hoffman JR, et al. **The Canadian C-spine rule.** *N Engl J Med* 2004;350:1467–69 CrossRef Medline
10. Sharp AL, Huang BZ, Tang T, et al. **Implementation of the Canadian CT head rule and its association with use of computed tomography among patients with head injury.** *Ann Emerg Med* 2018;71:54–63 CrossRef Medline
11. Mower WR, Gupta M, Rodriguez R, et al. **Validation of the sensitivity of the National Emergency X-Radiography Utilization Study (NEXUS) head computed tomographic (CT) decision instrument for selective imaging of blunt head injury patients: an observational study.** *PLoS Med* 2017;14:e1002313 CrossRef Medline
12. DSI Home. **FDA Cleared AI Algorithms. American College of Radiology.** 2020. https://models.acrdsi.org/. Accessed September 11, 2020
13. Benjamens S, Dhunnoo P, Mesko B. **The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database.** *NPJ Digit Med* 2020;3:118 CrossRef Medline
14. Alwosheel A, van Cranenburgh S, Chorus CG. **Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis.** *J Choice Model* 2018;28:167–82 CrossRef
15. Park SH, Han K. **Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction.** *Radiology* 2018;286:800–09 CrossRef Medline
16. Kim DW, Jang HY, Kim KW, et al. **Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers.** *Korean J Radiol* 2019;20:405–10 CrossRef Medline
17. Liu X, Faes L, Kale AU, et al. **A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis.** *Lancet Digit Health* 2019;1:e271–97 CrossRef Medline
18. Zech JR, Badgeley MA, Liu M, et al. **Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study.** *PLoS Med* 2018;15:e1002683 CrossRef Medline
19. U.S. Food and Drug Administration. **K190896.** 2019 https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190896.pdf. Accessed February 19, 2021
20. Burns JE, Yao J, Munoz H, et al. **Automated detection, localization, and classification of traumatic vertebral body fractures in the thoracic and lumbar spine at CT.** *Radiology* 2016;278:64–73 CrossRef Medline

21. Burns JE, Yao J, Summers RM. **Vertebral body compression fractures and bone density: automated detection and classification on CT images.** *Radiology* 2017;284:788–97 CrossRef Medline
22. Tomita N, Cheung YY, Hassanpour S. **Deep neural networks for automatic detection of osteoporotic vertebral fractures on CT scans.** *Comput Biol Med and Med* 2018;98:8–15 CrossRef Medline
23. Roth HR, Wang Y, Yao J, et al. **Deep convolutional networks for automated detection of posterior-element fractures on spine CT.** *Computer Science > Computer Vision and Pattern Recognition* January 29, 2016. https://arxiv.org/abs/1602.00020. Accessed September 2, 2020
24. Ginat DT. **Analysis of head CT scans flagged by deep learning software for acute intracranial hemorrhage.** *Neuroradiology* 2020;62:335–40 CrossRef Medline
25. Ojeda PZ, Zawaideh M, Mossa-Basha M, et al. **The utility of deep learning: evaluation of a convolutional neural network for detection of intracranial bleeds on non-contrast head computed tomography studies.** In: Proeedings of SPIE Medical Imaging, San Diego California. February 16–21, 2019