


# Deep Learning Assistance Closes the Accuracy Gap in Fracture Detection Across Clinician Types

Pamela G. Anderson PhD<sup>1</sup> , Graham L. Baum PhD<sup>1</sup>, Nora Keathley BS<sup>1</sup>, Serge Sicular MD<sup>1,2</sup>, Srivas Venkatesh MS<sup>1</sup>, Anuj Sharma MS<sup>1</sup>, Aaron Daluiski MD<sup>3</sup>, Hollis Potter MD<sup>3</sup>, Robert Hotchkiss MD<sup>3</sup>, Robert V. Lindsey PhD<sup>1</sup>, Rebecca M. Jones PhD<sup>1</sup>

Received: 4 January 2022 / Accepted: 5 August 2022 / Published online: 9 September 2022

Copyright © 2023 The Author(s). Published by Wolters Kluwer Health, Inc. on behalf of the Association of Bone and Joint Surgeons

## Abstract

**Background** Missed fractures are the most common diagnostic errors in musculoskeletal imaging and can result in treatment delays and preventable morbidity. Deep learning, a subfield of artificial intelligence, can be used to accurately detect fractures by training algorithms to emulate the judgments of expert clinicians. Deep learning systems that detect fractures are often limited to specific anatomic regions and require regulatory approval to be used in practice. Once these hurdles are overcome, deep

learning systems have the potential to improve clinician diagnostic accuracy and patient care.

**Questions/purposes** This study aimed to evaluate whether a Food and Drug Administration–cleared deep learning system that identifies fractures in adult musculoskeletal radiographs would improve diagnostic accuracy for fracture detection across different types of clinicians. Specifically, this study asked: (1) What are the trends in musculoskeletal radiograph interpretation by different clinician types in the publicly available Medicare claims data? (2) Does the deep learning system improve clinician accuracy in diagnosing fractures on radiographs and, if so, is there a greater benefit for clinicians with limited training in musculoskeletal imaging?

**Methods** We used the publicly available Medicare Part B Physician/Supplier Procedure Summary data provided by the Centers for Medicare & Medicaid Services to determine the trends in musculoskeletal radiograph interpretation by clinician type. In addition, we conducted a multiple-reader, multiple-case study to assess whether clinician accuracy in diagnosing fractures on radiographs was superior when aided by the deep learning system compared with when unaided. Twenty-four clinicians (radiologists, orthopaedic surgeons, physician assistants, primary care physicians, and emergency medicine physicians) with a median (range) of 16 years (2 to 37) of experience postresidency each assessed 175 unique musculoskeletal radiographic cases under aided and unaided conditions (4200 total case-physician pairs per condition). These cases were comprised of radiographs from 12 different anatomic regions (ankle, clavicle, elbow, femur, forearm, hip, humerus, knee, pelvis, shoulder, tibia and fibula, and wrist) and were randomly selected from 12 hospitals and healthcare centers.

This study was funded by Imagen Technologies.

One or more authors (PGA, SS, SV, RVL, RMJ) are employees and equity holders at Imagen Technologies. One or more authors (GLB, NK) are employees of Imagen Technologies. One or more authors (AS, AD, HP, RH) are equity holders at Imagen Technologies.

All ICMJE Conflict of Interest Forms for authors and *Clinical Orthopaedics and Related Research*® editors and board members are on file with the publication and can be viewed on request.

*Clinical Orthopaedics and Related Research*® neither advocates nor endorses the use of any treatment, drug, or device. Readers are encouraged to always seek additional information, including FDA approval status, of any drug or device before clinical use.


Ethical approval for this study was obtained from the New England Independent Review Board (number 120190100).

This work was performed at Imagen Technologies, New York, NY, USA.

<sup>1</sup>Imagen Technologies, New York, NY, USA

<sup>2</sup>The Mount Sinai Hospital, New York, NY, USA

<sup>3</sup>Hospital for Special Surgery, New York, NY, USA

P. G. Anderson , Imagen Technologies, 594 Broadway #701, New York, NY 10012, USA, Email: pami.anderson@imagen.ai

The gold standard for fracture diagnosis was the majority opinion of three US board-certified orthopaedic surgeons or radiologists who independently interpreted the case. The clinicians' diagnostic accuracy was determined by the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, sensitivity, and specificity. Secondary analyses evaluated the fracture miss rate (1-sensitivity) by clinicians with and without extensive training in musculoskeletal imaging.

**Results** Medicare claims data revealed that physician assistants showed the greatest increase in interpretation of musculoskeletal radiographs within the analyzed time period (2012 to 2018), although clinicians with extensive training in imaging (radiologists and orthopaedic surgeons) still interpreted the majority of the musculoskeletal radiographs. Clinicians aided by the deep learning system had higher accuracy diagnosing fractures in radiographs compared with when unaided (unaided AUC: 0.90 [95% CI 0.89 to 0.92]; aided AUC: 0.94 [95% CI 0.93 to 0.95]; difference in least square mean per the Dorfman, Berbaum, Metz model AUC: 0.04 [95% CI 0.01 to 0.07];  $p < 0.01$ ). Clinician sensitivity increased when aided compared with when unaided (aided: 90% [95% CI 88% to 92%]; unaided: 82% [95% CI 79% to 84%]), and specificity increased when aided compared with when unaided (aided: 92% [95% CI 91% to 93%]; unaided: 89% [95% CI 88% to 90%]). Clinicians with limited training in musculoskeletal imaging missed a higher percentage of fractures when unaided compared with radiologists (miss rate for clinicians with limited imaging training: 20% [95% CI 17% to 24%]; miss rate for radiologists: 14% [95% CI 9% to 19%]). However, when assisted by the deep learning system, clinicians with limited training in musculoskeletal imaging reduced their fracture miss rate, resulting in a similar miss rate to radiologists (miss rate for clinicians with limited imaging training: 9% [95% CI 7% to 12%]; miss rate for radiologists: 10% [95% CI 6% to 15%]).

**Conclusion** Clinicians were more accurate at diagnosing fractures when aided by the deep learning system, particularly those clinicians with limited training in musculoskeletal image interpretation. Reducing the number of missed fractures may allow for improved patient care and increased patient mobility.

**Level of Evidence** Level III, diagnostic study.

## Introduction

Missed fractures are the most common diagnostic errors made by clinicians interpreting musculoskeletal radiographs and cause treatment delays, unnecessary medical costs, malpractice lawsuits, and preventable morbidity [20, 33]. Concurrent increases in musculoskeletal fractures and radiography use rates over the past 20 years have resulted

in excessive workloads for clinicians interpreting radiographs, which can cause fatigue and increase susceptibility to interpretational errors [2, 3, 19, 20, 31, 33]. Radiographic interpretation has also increasingly been performed by nonradiologists with limited training in musculoskeletal imaging [3, 31, 37], who are more prone to diagnostic errors [32, 23]. Nonradiologists often perform an initial radiograph interpretation when treating patients in settings such as the emergency department or outpatient clinics, and discrepancies were found between the initial radiograph interpretations from nonradiologists and the radiologists' final read [13, 18]. Developing tools that can reduce the gap in diagnostic accuracy between clinicians with and without extensive training in interpreting radiographs could improve patient outcomes and reduce medical costs associated with missed fractures, particularly in the Medicare-age population (age 65 and older) [4, 26, 28]. Fractures are the most common musculoskeletal condition resulting in hospitalization among Medicare enrollees [34]. Studies have shown that fractures in older patients can result in increased mortality, reduced mobility, and greater difficulties with living independently [10, 22, 30].

Deep learning, a subfield of artificial intelligence, can be used to accurately detect fractures by training algorithms to emulate the judgments of expert clinicians [6, 24, 29, 35]. Deep learning systems for fracture detection and localization are often limited in scope to specific anatomic regions and clinical settings. Studies on deep learning systems for fracture detection have reported standalone performance [7, 8, 24]; however, existing studies evaluating whether deep learning systems improve the diagnostic accuracy of clinicians [9, 29] have not accounted for the wide range of experience and specialization across clinicians who interpret radiographs. Additionally, these deep learning systems have not been cleared by the United States Food and Drug Administration (FDA) as safe and effective medical devices, which is crucial for widespread use and adoption in daily clinical practice [9, 25, 35, 40].

In this study, we investigated the trends of musculoskeletal radiograph interpretation by clinician type from the publicly available Centers for Medicare & Medicaid Services (CMS) database to understand the potential impact of a deep learning system to assist different clinician types. It is critical to assess whether FractureDetect (Imagen Technologies Inc), a deep learning system for fracture detection, benefits clinicians with extensive training in musculoskeletal imaging and those with limited training in musculoskeletal imaging. Clinicians with limited training in musculoskeletal imaging are typically the first to evaluate a patient after trauma and are increasingly performing radiographic interpretation and imaging-guided treatments [27, 31, 37]. Therefore, there could be a significant positive impact on patients if the deep learning system improves the abilities of clinicians with

limited training in musculoskeletal imaging to diagnose fractures.

The study addressed the following questions: (1) What are the trends in musculoskeletal radiograph interpretation by different clinician types in the publicly available Medicare claims data? (2) Does the deep learning system improve clinician accuracy in diagnosing fractures on radiographs and, if so, is there a greater benefit for clinicians with limited training in musculoskeletal imaging?

## Materials and Methods

### *Volume of Musculoskeletal Radiographs for Medicare Beneficiaries*

We report the volume of musculoskeletal radiographs taken in Medicare beneficiaries between 2012 and 2018 to determine which clinician types interpreted musculoskeletal radiographs. We used the publicly available Medicare Part B Physician/Supplier Procedure Summary (PSPS) data provided by the CMS. The CMS data encompassed all medical services to beneficiaries in the traditional fee-for-service population and included the volume of claims billed by clinician type for all Current Procedural Terminology (CPT) codes. The CPT codes were filtered to include only musculoskeletal radiographs corresponding to the 12 anatomic regions (ankle, clavicle, elbow, femur, forearm, hip, humerus, knee, pelvis, shoulder, tibia and fibula, and wrist) indicated for use by the deep learning system [15]. We evaluated the volume of radiograph interpretation by clinician type in this study, which included radiologists, orthopaedic surgeons, physician assistants, primary care physicians, and emergency medicine physicians. The primary care physicians included physicians with specialties in family practice and internal medicine. Radiograph use per 1000 beneficiaries was calculated for each clinician type by taking the sum of radiograph claim volume and dividing by the number of beneficiaries per year [5]. Because PSPS data were tabulated based on global plus professional component claims (excluding technical component-only claims), use rates reflect the volume of radiograph interpretation for each clinician type [31]. We also examined care settings (such as office or hospital) using the place of service codes in the 2018 PSPS data to determine the setting in which different clinician types interpret musculoskeletal radiographs.

### *Clinical Study Design and Setting*

This retrospective clinical study followed FDA guidance [41] to evaluate the performance of the deep learning system. We

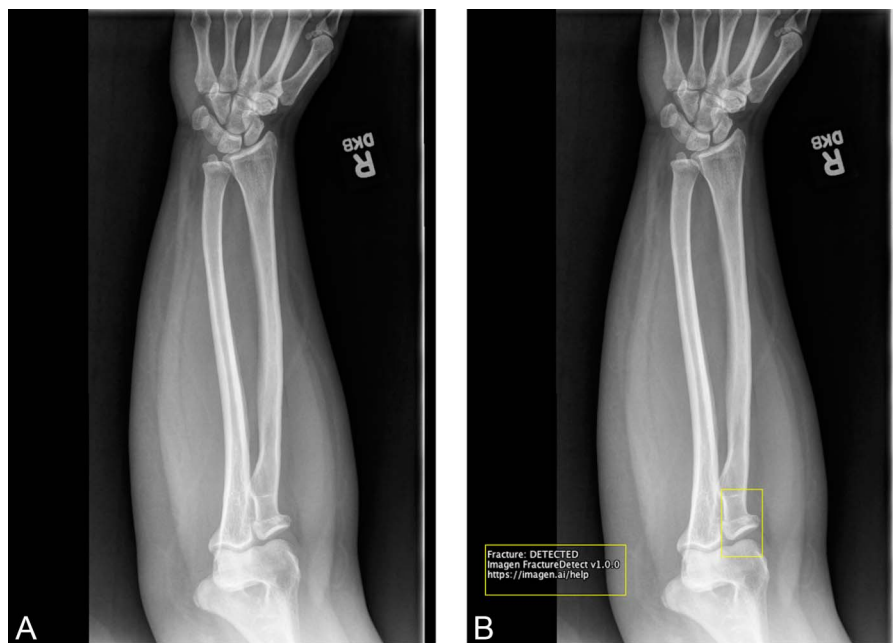
randomly sampled deidentified cases collected from 12 hospitals and healthcare centers to have a representative set of patients. The sampling process was designed so that there was a relatively balanced number of cases across anatomic regions (Supplementary Table 1; <http://links.lww.com/CORR/A933>). All patients were adults at least 22 years old (Supplementary Table 2; <http://links.lww.com/CORR/A934>). No radiographs used in the development of the deep learning system or standalone testing [24] were present in the clinical testing dataset. There were 175 patient cases within the deep learning system's indications for use [15], and the results from these patients are reported in this article. There were 67 cases from patients aged 65 and older, and these cases were additionally analyzed to examine the impact of the deep learning system on the Medicare-age population (Supplementary Digital Content 1; <http://links.lww.com/CORR/A935>). A case consisted of radiographs from a single patient's study without any clinical history provided. Based on a power analysis and an assumed difference in aided versus unaided areas under the curve (AUCs) of 0.04 derived from prior work [14], the study design using 24 clinicians and 175 cases provided more than 90% power.

Twenty-four clinicians interpreted cases in the study. Clinicians had a median (range) of 16 years (2 to 37) of experience. Four radiologists and four orthopaedic surgeons were included in the study and represent clinician types with extensive training in musculoskeletal imaging. Four internal medicine physicians, four family medicine physicians, four emergency medicine physicians, and four emergency medicine physician assistants (referred to as "physician assistants") were included in the study as clinicians with limited training in musculoskeletal imaging. The intended users of the deep learning system are clinicians of various specialties, not only radiologists; therefore, an equal number of clinicians in these six specialties were selected to properly evaluate the impact of the deep learning system on performance across a range of clinician types.

The study consisted of two independent reading sessions separated by a washout period of at least 28 days. Clinicians interpreted all cases twice. In the first session, half the cases were aided by the deep learning system and the other half were unaided. In the second session, all cases were read in the opposite condition. Cases were assigned using randomized stratification to reduce case order effects. Each clinician was asked to determine the presence or absence of a fracture in each case and provide a confidence score (0 to 100) of their assessment.

### *Gold Standard for Fracture Diagnosis*

The gold standard for diagnosis (also known as the "ground truth label") of a fracture in this study was the



**Fig. 1 A-B** This is an example radiograph from the clinical study with and without the deep learning system overlay. **(A)** The radiograph of a fracture in the elbow without the deep learning system overlay (the unaided condition). **(B)** The radiograph with the deep learning system toggleable overlay (the aided condition). The overlay consists of a bounding box surrounding the site of the fracture in the radiograph (radial neck) and a text box in the bottom left corner of the radiograph stating “Fracture: DETECTED”.

majority opinion of three US board-certified orthopaedic surgeons or radiologists who independently interpreted the case. These orthopaedic surgeons and radiologists had a median (range) of 13 years (4 to 35) of experience postresidency. If any image in a case had a fracture-positive ground truth label, the case was deemed as having a fracture. Otherwise, the case was deemed as not having a fracture. Fractures were identified in 24% (42 of 175) of cases. There was complete agreement among the three physicians providing ground truth labels for 87% (153 of 175) of cases.

#### Deep Learning System

The deep learning system served as a concurrent reading aid for clinicians in detecting and diagnosing fractures across 12 anatomic regions of the musculoskeletal system (Supplementary Digital Content 1; <http://links.lww.com/CORR/A935>). The deep learning system produces a binary text output representing the determination of whether any fractures are visible on a radiograph and a set of bounding boxes surrounding fracture sites (Fig. 1). Further details of the deep learning system and its standalone performance have been described elsewhere

[24]. The clinical study on the deep learning system described in this article, in part, led to the FDA’s clearance of the device to assist clinicians in detecting fractures on musculoskeletal radiographs [15].

#### Ethical Approval

The study was Health Insurance Portability and Accounting Act-compliant and was approved by the New England Independent Review Board.

#### Statistical Analysis

We calculated the AUC of the receiver operating characteristic (ROC) curve to evaluate clinician accuracy. The AUC was determined by finding the area under the ROC curve. The ROC curve can be found by transforming the readers’ confidence scores into binary responses based off of stepwise threshold values (for example, if a confidence score was below a threshold, then the response was considered to be “fracture absent” versus if a confidence score was above a threshold, then the response was considered to be “fracture present”). At each step of the confidence score

**Table 1.** Percentage of musculoskeletal radiograph interpretation and volume per 1000 beneficiaries by year by clinician type in Medicare claims data for the 12 anatomical regions indicated for use by the deep learning system

Clinician type <sup>a</sup>	2012 (n = 504)	2015 (n = 516)	2018 (n = 484)	% change in volume 2012 to 2018
Radiologists	52 (263)	53 (273)	54 (262)	-1
Orthopaedic surgeons	33 (167)	32 (165)	30 (146)	-13
Physician assistants	2 (11)	3 (17)	4 (22)	89
Primary care physicians	1 (4)	1 (4)	1 (3)	-39
Emergency medicine physicians	1 (3)	1 (3)	1 (2)	-22

Data presented as % (n).

<sup>a</sup>The percentage of the total radiograph interpretation volume by each clinician type does not sum to 100% due to the exclusion of clinician types that interpret radiographs but were not included in the clinical study (such as, podiatrists and hand surgeons).

threshold value, the sensitivity and specificity were calculated and were used to plot the ROC curve (sensitivity versus 1-specificity). If clinicians were to guess at random whether there were fractures, the AUC would be 0.5, and if clinicians were perfect at identifying fractures (relative to the gold standard), the AUC would be 1.0. We used the Dorfman, Berbaum, Metz model as the primary endpoint of this clinical study to evaluate whether there was a statistical improvement between an AUC calculated in one condition (such as, clinicians aided by the deep learning system) compared with another condition (such as, unaided clinicians) [11, 21]. We also calculated the sensitivity, specificity, miss rate (1-sensitivity), positive predictive value, and negative predictive value, across all clinicians and for each clinician type by treating each case, read by each clinician, independently. To parallel the clinician types in the Medicare data, we combined internal medicine and family medicine physicians into a primary care physician group. To examine the deep learning system's performance on the Medicare-age population, we calculated the miss rate for clinicians with and without extensive training in musculoskeletal imaging for all images acquired for patients at least 65 years old (n = 67). All statistical analyses were performed with R (version 3.6.1) and Python (version 3.6.4).

We performed a series of analyses to evaluate the potential impact of the deep learning system on the United States Medicare-age population (Supplementary Digital Content 1; <http://links.lww.com/CORR/A935>).

## Results

### *Trends in Musculoskeletal Interpretation by Clinician Type*

The Medicare claims data demonstrated that physician assistants showed the greatest increase (89%) in interpretation of musculoskeletal radiographs from 2012 to 2018 (Table 1).

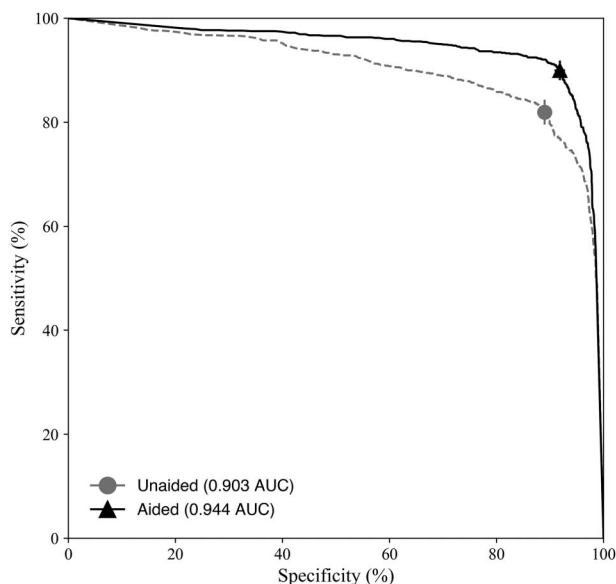
In 2018, physician assistants provided 4% of radiograph interpretations (Supplementary Table 3; <http://links.lww.com/CORR/A936>), primarily interpreting radiographs in office and urgent care settings (Supplementary Table 4; <http://links.lww.com/CORR/A937>). As expected, clinicians with extensive training in musculoskeletal imaging (radiologists and orthopaedic surgeons) interpreted most of the musculoskeletal radiographs (84% total volume) in 2018, with a small decrease in the number of interpretations from 2012 to 2018 (Table 1). In 2018, radiologists performed most of the radiograph interpretations in hospital settings. Specifically, radiologists performed 97% and 98% of all radiographic interpretations in emergency rooms and inpatient hospitals, respectively (Supplementary Table 4; <http://links.lww.com/CORR/A937>).

### *Deep Learning System Improved Diagnostic Accuracy of Fractures on Radiographs and Benefited Clinicians With Limited Training in Musculoskeletal Imaging*

The deep learning system improved clinicians' accuracy at diagnosing musculoskeletal fractures (unaided AUC: 0.903 [95% CI 0.890 to 0.916]; aided AUC: 0.944 [95% CI 0.934 to 0.954]; difference in least square mean per the Dorfman, Berbaum, Metz model AUC: 0.041 [95% CI 0.013 to 0.069];  $p < 0.01$ ). Clinicians demonstrated improvements in sensitivity and specificity when assisted by the deep learning system (Fig. 2). Clinician sensitivity increased when aided compared with when unaided (aided: 90% [95% CI 88% to 92%]; unaided: 82% [95% CI 79% to 84%]), and specificity increased when aided compared with when unaided (aided: 92% [95% CI 91% to 93%]; unaided: 89% [95% CI 88% to 90%]).

Sensitivity and specificity improved when aided by the deep learning system for different clinician types (Fig. 3). Clinicians with limited training in interpreting musculoskeletal radiographs (primary care physicians, physician assistants, and emergency medicine physicians) showed





**Fig. 2** The deep learning system increased clinician diagnostic accuracy at detecting musculoskeletal fractures, as demonstrated by ROC curves, sensitivity, and specificity for clinicians unaided and aided by the deep learning system. Error bars represent the 95% bootstrap CIs (m = 1000).

the greatest improvements in sensitivity and specificity when assisted by the deep learning system. Clinicians with limited training in musculoskeletal imaging missed a higher percentage of fractures when unaided compared with radiologists (miss rate for clinicians with limited imaging training: 20% [95% CI 17% to 24%]; miss rate for radiologists: 14% [95% CI 9% to 19%]). However, when assisted by the deep learning system, clinicians with limited training in musculoskeletal imaging reduced their fracture miss rate, resulting in a similar miss rate to radiologists (miss rate for clinicians with limited imaging training: 9% [95% CI 7% to 12%]; miss rate for radiologists: 10% [95% CI 6% to 15%]) (Supplementary Table 5; <http://links.lww.com/CORR/A938>). The sensitivity, miss rate, specificity, positive predictive value, and negative predictive value for all clinicians and per clinician type were measured (Supplementary Table 5; <http://links.lww.com/CORR/A938>), in addition to the average confidence scores of the clinicians' assessments (Supplementary Table 6; <http://links.lww.com/CORR/A939>).

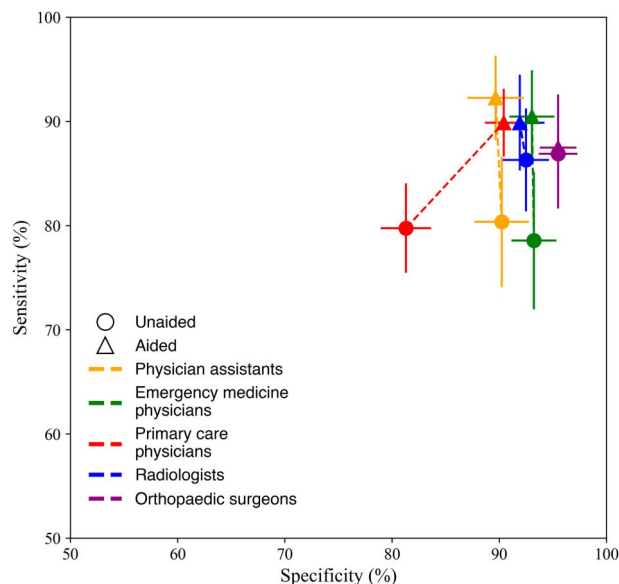
In Medicare-age cases, a population with relatively high fracture prevalence, clinicians with limited training and clinicians with extensive training in musculoskeletal imaging both missed fewer fractures when assisted by the deep learning system (Fig. 4). Clinicians with limited training in musculoskeletal imaging had a 61% relative reduction in missed fractures when aided by the deep learning system. Clinicians with extensive training in musculoskeletal imaging (radiologists and orthopaedic

surgeons) had a 38% relative reduction in missed fractures when assisted by the deep learning system. For each clinician type, the fracture miss rate decreased when aided by the deep learning system (Supplementary Fig. 1; <http://links.lww.com/CORR/A940>).

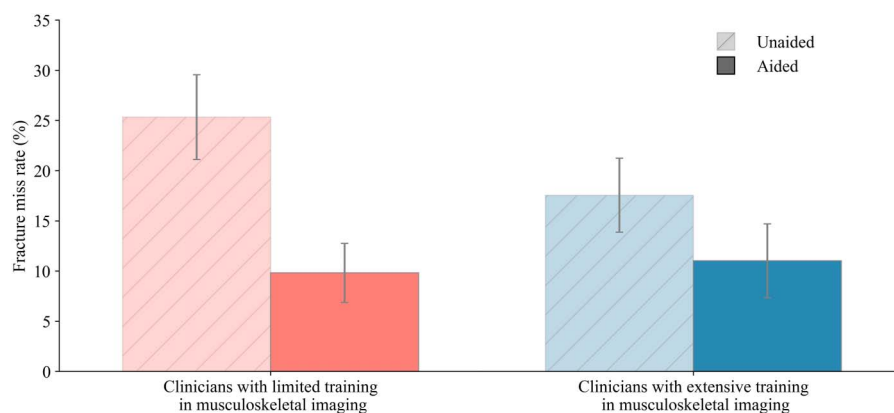
Our results demonstrate that the deep learning system has the potential to benefit the United States Medicare-age population by reducing the number of missed fractures in musculoskeletal radiographs (Supplementary Table 7; <http://links.lww.com/CORR/A941>).

**Discussion**

The interpretation of radiographs is challenging, and diagnostic errors often occur in busy clinical settings that rely on overburdened clinicians who lack subspecialized expertise. One solution is to assist clinicians' radiograph interpretation with a deep learning system. Before deploying in the clinical setting, it is necessary to evaluate whether a deep learning system helps different types of clinicians diagnose fractures in radiographs more accurately. This clinical study demonstrated that FractureDetect, the first FDA-cleared deep learning system trained to detect and localize musculoskeletal fractures in multiple anatomic regions, improved the diagnostic accuracy of many types of clinicians.



**Fig. 3** The aided and unaided performance for detecting musculoskeletal fractures is different by clinician type. Clinicians with limited training in musculoskeletal imaging (physician assistants, primary care physicians, and emergency medicine physicians) had the largest increase in sensitivity and specificity when aided by the deep learning system. Error bars represent the 95% bootstrap CIs (m = 1000). A color image accompanies the online version of this article.



**Fig. 4** The deep learning system reduced missed fractures across the Medicare-age population for clinicians with limited training in musculoskeletal imaging (physician assistants, primary care physicians, and emergency medicine physicians) and clinicians with extensive training in musculoskeletal imaging (radiologists, orthopaedic surgeons). Error bars represent the 95% bootstrap confidence intervals ( $m = 1000$ ).

### Limitations

There are notable limitations of this study. First, the study design involved instructing clinicians to only identify fractures on musculoskeletal radiographs and not to comment on other possible abnormalities, which could inflate clinician performance for fracture detection. Although instructing clinicians to focus on identifying fractures could increase diagnostic accuracy for fracture detection, it would equally impact clinician accuracy in both unaided and aided experimental conditions and thus likely had a negligible impact on the main conclusions of the study. Second, although it was expected that the deep learning system would benefit clinicians with limited training in musculoskeletal imaging more than orthopaedic surgeons or radiologists, the study was not powered to test for statistical differences between the two groups of clinicians. Despite the small sample size, the data showed a clear enhanced benefit for clinicians with limited training in musculoskeletal imaging relative to clinicians with extensive training. Third, although the Dorfman, Berbaum, Metz model is the suggested statistical method to evaluate multiple reader, multiple case studies [41], there are limitations to this analytical approach and alternative methods have been proposed [38].

There are also multiple limitations of this study that could be addressed with a prospective clinical study. For example, this study did not incorporate clinical history as the readers were interpreting radiographs, did not evaluate patient outcomes as a result of a missed fracture, and did not assess whether fractures required clinical evaluation. However, studies have shown that minor fractures, including those that require no treatment, are important to identify to prevent possible adverse patient outcomes and to minimize complications that can arise from fractures [16, 17, 36]. Future studies that evaluate patients in a

prospective setting could determine the impact of the deep learning system on patient outcomes.

### Trends in Musculoskeletal Interpretation by Clinician Type

The analysis of Medicare radiograph interpretation volumes between 2012 and 2018 revealed a large increase in the proportion of physician assistant interpretations relative to radiologists and other clinicians. These results are consistent with previous studies showing that nonradiologists are increasingly performing diagnostic imaging and imaging-guided treatments [27, 31, 37]. The care setting analysis suggested that physician assistant interpretation predominantly occurred in office and urgent care settings and that radiologists continued to perform most hospital interpretations. Although it is beneficial to increase patient access to care outside the hospital setting, it is critical that this shift toward radiograph interpretation by clinicians with limited training in musculoskeletal imaging does not lead to an increase in costly diagnostic errors. The deep learning system provides a scalable solution for improving the standard of care across different clinical settings.

### Deep Learning System Improved Diagnostic Accuracy of Fractures on Radiographs and Benefited Clinicians With Limited Training in Musculoskeletal Imaging

The deep learning system improved diagnostic accuracy for clinicians both with and without extensive training in musculoskeletal imaging. Clinicians with limited training in musculoskeletal imaging had the largest increase in

diagnostic accuracy when aided by the deep learning system. Previous studies have evaluated the efficacy of deep learning systems for fracture detection on radiographs with only two clinician specialties (radiology and emergency medicine) [12, 29]; the current study broadened the clinician sample by including five specialties. Specifically, the deep learning system enabled physician assistants, primary care physicians, and emergency medicine physicians to detect fractures with diagnostic accuracy comparable with that of radiologists and orthopaedic surgeons. Although most musculoskeletal radiographs are still interpreted by radiologists and orthopaedic surgeons, our study shows that the deep learning system improves the performance of multiple clinician types who are interpreting radiographs, especially those who may benefit even more from assistance. These clinicians with limited training in musculoskeletal imaging are often the first to evaluate a patient after trauma; therefore, prompt, accurate diagnosis is important in ensuring effective treatment and mitigating the risk of patient complications.

We also demonstrated that the deep learning system can improve diagnostic accuracy and reduce the rate of missed fractures in Medicare-age patients. We integrated the analysis of Medicare claims data with published fracture rates [28] and clinician miss rates from our clinical study to estimate the number of missed fractures per year (Supplementary Digital Content 1; <http://links.lww.com/CORR/A935>). We estimated that when aided by the deep learning system, clinicians could miss 43% (52,346 of 121,393) fewer fractures per year across the Medicare-age population (Supplementary Table 7; <http://links.lww.com/CORR/A941>). Given the high direct and indirect costs of missed fractures [1, 4, 39], these results highlight the deep learning system's potential impact on the US medical system. We were unable to estimate the fracture miss rate per anatomic region because we had a limited number of cases per anatomic region. Therefore, the overall miss rate may be over- or underestimated, which would affect our reported number of fractures missed in the Medicare-age population. Future studies that are able to use a precise fracture rate for each of the 12 anatomic regions may result in a better estimate for how the deep learning system will improve clinicians' ability to diagnose fractures in the Medicare-age population.

### Conclusion

The FDA-cleared deep learning system [15] reduced diagnostic errors in fracture detection and enabled clinicians with limited training in musculoskeletal imaging to have similar performance to radiologists. The deep learning system has the potential to reduce the number of missed

fractures and improve patient care, especially in the Medicare-age population.

This is an open-access article distributed under the terms of the [Creative Commons Attribution-Non Commercial-No Derivatives License 4.0 \(CCBY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/), where it is permissible to download and share the work provided it is properly cited. The work cannot be changed in any way or used commercially without permission from the journal.

**Acknowledgment** We thank Tung Phan PhD for providing valuable statistical advice.

### References

1. Adeyemi A, Delhougne G. Incidence and economic burden of intertrochanteric fracture: a Medicare claims database analysis. *JB JS Open Access*. 2019;4:e0045.
2. Amin S, Achenbach SJ, Atkinson EJ, Khosla S, Melton LJ. Trends in fracture incidence: a population-based study over 20 years. *J Bone Miner Res*. 2014;29:581-589.
3. Berlin L. Defending the "missed" radiographic diagnosis. *AJR Am J Roentgenol*. 2001;176:317-322.
4. Burge R, Dawson-Hughes B, Solomon DH, Wong JB, King A, Tosteson A. Incidence and economic burden of osteoporosis-related fractures in the United States, 2005-2025. *J Bone Miner Res*. 2007;22:465-475.
5. Centers for Medicare and Medicaid Services. MDCR ENROLL AB 9. Original Medicare enrollment: part A and/or part B total, aged, and disabled enrollees, calendar years 2013-2018. Available at: <https://www.cms.gov/files/document/2018-mdcr-enroll-ab-9.pdf>. Accessed June 10, 2021.
6. Chea P, Mandell JC. Current applications and future directions of deep learning in musculoskeletal radiology. *Skeletal Radiol*. 2020;49:183-197.
7. Chen HY, Hsu BWY, Yin Y-K, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. *PLoS One*. 2021;16:e0245992.
8. Cheng C-T, Wang Y, Chen H-W, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. *Nat Commun*. 2021;12:1-10.
9. Chung SW, Han SS, Lee JW, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop*. 2018;89:468-473.
10. Clement ND, Aitken SA, Duckworth AD, McQueen MM, Court-Brown CM. The outcome of fractures in very elderly patients. *J Bone Joint Surg Br*. 2011;96:806-810.
11. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest Radiol*. 1992;27:723-731.
12. Duron L, Ducarouge A, Gillibert A, et al. Assessment of an AI aid in detection of adult appendicular skeletal fractures by emergency physicians and radiologists: a multicenter cross-sectional diagnostic study. *Radiology*. 2021;300:120-129.
13. Fernholm R, Pukk Härenstam K, Wachtler C, Nilsson GH, Holzmann MJ, Carlsson AC. Diagnostic errors reported in primary healthcare and emergency departments: a retrospective and descriptive cohort study of 4830 reported cases of preventable harm in Sweden. *Eur J Gen Pract*. 2019;25:128-135.
14. Food and Drug Administration Center for Devices and Radiological Health. Evaluation of automatic Class III designation for OsteoDetect DEN180005. <https://www.accessdata.fda.gov>



- [gov/cdrh\\_docs/reviews/DEN180005.pdf](https://www.accessdata.fda.gov/cdrh_docs/reviews/DEN180005.pdf). Accessed June 10, 2021.
15. Food and Drug Administration Center for Devices and Radiological Health. 510(k) Premarket Notification FractureDetect (FX) K193417. Available at: [https://www.accessdata.fda.gov/cdrh\\_docs/pdf19/K193417.pdf](https://www.accessdata.fda.gov/cdrh_docs/pdf19/K193417.pdf). Accessed June 10, 2021.
  16. Gardner MJ, Demetropoulos D, Shindle MK, Griffith MH, Lane JM. Osteoporosis and skeletal fractures. *HSS Journal*. 2006;2:62-69.
  17. Gillespie S, Cowell F, Cheung G, Brown D. Can we reduce the incidence of complex regional pain syndrome type I in distal radius fractures? The Liverpool experience. *Hand Therapy*. 2016;21:123-130.
  18. Guerhazi A, Tannoury C, Kompel AJ, et al. Improving radiographic fracture recognition performance and efficiency using artificial intelligence. *Radiology*. 2022;302:627-636.
  19. Guly H. Diagnostic errors in an accident and emergency department. *Emerg Med J*. 2001;18:263-269.
  20. Hallas P, Ellingsen T. Errors in fracture diagnoses in the emergency department – characteristics of patients and diurnal variation. *BMC Emerg Med*. 2006;6:1-5.
  21. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. *Acad Radiol*. 2008;15:647-661.
  22. Holt G, Smith R, Duncan K, Hutchison JD, Gregori A. Outcome after surgery for the treatment of hip fracture in the extremely elderly. *J Bone Joint Surg Am*. 2008;90:1899-1905.
  23. Hussain F, Cooper A, Carson-Stevens A, et al. Diagnostic error in the emergency department: learning from national patient safety incident report analysis. *BMC Emerg Med*. 2019;19:1-9.
  24. Jones RM, Sharma A, Hotchkiss R, et al. Assessment of a deep-learning system for fracture detection in musculoskeletal radiographs. *NPJ Digit Med*. 2020;3:1-6.
  25. Krogue JD, Cheng KV, Hwang KM, et al. Automatic hip fracture identification and functional subclassification with deep learning. *Radiol Artif Intell*. 2020;2:e190023.
  26. Lau E, Ong K, Kurtz S, Schmier J, Edidin A. Mortality following the diagnosis of a vertebral compression fracture in the Medicare population. *J Bone Joint Surg Am*. 2008;90:1479-1486.
  27. Levin DC, Rao VM, Parker L, Frangos AJ, Sunshine JH. Medicare payments for noninvasive diagnostic imaging are now higher to nonradiologist physicians than to radiologists. *J Am Coll Radiol*. 2011;8:26-32.
  28. Lewiecki EM, Chastek B, Sundquist K, et al. Osteoporotic fracture trends in a population of US managed care enrollees from 2007 to 2017. *Osteoporos Int*. 2020;31:1299-1304.
  29. Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci USA*. 2018;115:11591-11596.
  30. Matityahu A, Elson J, Morshed S, Marmor M. Survivorship and severe complications are worse for octogenarians and elderly patients with pelvis fractures as compared to adults: data from the national trauma data bank. *J Osteoporos*. 2012;2012:475739.
  31. Mizrahi DJ, Parker L, Zoga AM, Levin DC. National trends in the utilization of skeletal radiography from 2003 to 2015. *J Am Coll Radiol*. 2018;15:1408-1414.
  32. McLauchlan CA, Jones K, Guly HR. Interpretation of trauma radiographs by junior doctors in accident and emergency departments: a cause for concern? *J Accid Emerg Med*. 1997;14:295-298.
  33. Moonen P-J, Mercelina L, Boer W, Fret T. Diagnostic error in the emergency department: follow up of patients with minor trauma in the outpatient clinic. *Scand J Trauma Resusc Emerg Med*. 2017;25:1-7.
  34. Office of the Surgeon General. *Bone Health and Osteoporosis: A Report of the Surgeon General*. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK45515/>. Accessed June 10, 2021.
  35. Olczak J, Fahlberg N, Maki A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. *Acta Orthop*. 2017;88:581-586.
  36. Rosen T, Bloemen EM, Harpe J, et al. Radiologists' training, experience, and attitudes about elder abuse detection. *AJR Am J Roentgenol*. 2016;207:1210.
  37. Rosman DA, Nsiah E, Hughes DR, Duszak R. Regional variation in Medicare payments for medical imaging: radiologists versus nonradiologists. *AJR Am J Roentgenol*. 2015;204:1042-1048.
  38. Song X, Xiao-Hua Z. A marginal model approach for analysis of multi-reader multi-test receiver operating characteristic (ROC) data. *Biostatistics*. 2005;6:303-312.
  39. Tran O, Silverman S, Xu X, et al. Long-term direct and indirect economic burden associated with osteoporotic fracture in US postmenopausal women. *Osteoporos Int*. 2021;32:1195-1205.
  40. Urakawa T, Tanaka Y, Goto S, Matsuzawa H, Watanabe K, Endo N. Detecting intertrochanteric hip fractures with orthopedist-level accuracy using a deep convolutional neural network. *Skeletal Radiol*. 2019;48:239-244.
  41. U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health. Clinical performance assessment: considerations for computer-assisted detection devices applied to radiology images and radiology device data in premarket notification (510(k)) submissions guidance for industry and FDA staff. Available at: <https://www.fda.gov/media/77642/download>. Accessed June 10, 2021.