Trauma Surgery & Acute Care Open

Practical computer vision application to detect hip fractures on pelvic X-rays: a bi-institutional study

Jeff Choi ,[1] James Z Hui,[2] David Spain,[3] Yi-Siang Su,[4] Chi-Tung Cheng ,[4] Chien-Hung Liao[4]

[1]General Surgery, Stanford University, Stanford, California, USA
[2]Radiology, Stanford University, Stanford, California, USA
[3]Surgery, Stanford University, Stanford, California, USA
[4]Trauma and Emergency Surgery, Chang Gung Memorial Hospital, Taoyuan, Taiwan

Correspondence to
Dr Jeff Choi; jc2226@stanford.edu

ABSTRACT

Background Pelvic X-ray (PXR) is a ubiquitous modality to diagnose hip fractures. However, not all healthcare settings employ round-the-clock radiologists and PXR sensitivity for diagnosing hip fracture may vary depending on digital display. We aimed to validate a computer vision algorithm to detect hip fractures across two institutions' heterogeneous patient populations. We hypothesized a convolutional neural network algorithm can accurately diagnose hip fractures on PXR and a web application can facilitate its bedside adoption.

Methods The development cohort comprised 4235 PXRs from Chang Gung Memorial Hospital (CGMH). The validation cohort comprised 500 randomly sampled PXRs from CGMH and Stanford's level I trauma centers. Xception was our convolutional neural network structure. We randomly applied image augmentation methods during training to account for image variations and used gradient-weighted class activation mapping to overlay heatmaps highlighting suspected fracture locations.

Results Our hip fracture detection algorithm's area under the receiver operating characteristic curves were 0.98 and 0.97 for CGMH and Stanford's validation cohorts, respectively. Besides negative predictive value (0.88 Stanford cohort), all performance metrics—sensitivity, specificity, predictive values, accuracy, and F1 score—were above 0.90 for both validation cohorts. Our web application allows users to upload PXR in multiple formats from desktops or mobile phones and displays probability of the image containing a hip fracture with heatmap localization of the suspected fracture location.

Discussion We refined and validated a high-performing computer vision algorithm to detect hip fractures on PXR. A web application facilitates algorithm use at the bedside, but the benefit of using our algorithm to supplement decision-making is likely institution dependent. Further study is required to confirm clinical validity and assess clinical utility of our algorithm.

Level of evidence III, Diagnostic tests or criteria.

INTRODUCTION

Hip fractures pose a considerable mortality and morbidity burden globally.[1 2] Long-term disability, increased risk for other adverse health conditions (eg, cardiovascular disease), and costly healthcare utilization are well-known sequelae.[3 4] Rapid hip fracture diagnosis is critical to mitigate both short and long-term adverse events. For operative candidates, guidelines recommend surgery for hip fractures be performed within 48 hours of injury.[5]

Pelvic X-ray (PXR) is an essential and ubiquitous modality to diagnose hip fractures. However, not all healthcare settings employ round-the-clock radiologists to interpret challenging plain radiographs. Moreover, PXR sensitivity for diagnosing hip fractures may vary depending on digital display and has been reported to be as low as 31% in a contemporary multi-institutional study.[6 7] Deep neural network learning is increasingly used to assist radiographic diagnoses, but limited data and lack of cross-institutional validation have hindered application for patients with hip fractures. A preliminary computer vision-based deep learning algorithm achieved 98% sensitivity in identifying hip fractures on PXR but has not been validated beyond a single-institution and single-race population.[8] Establishing clinical validity requires algorithm validation and refinement across heterogeneous populations.

We aimed to (A) refine and validate a computer vision algorithm to detect hip fractures across two institutions' heterogeneous patient populations, and (B) design a practical tool for bedside use. We hypothesized that a convolutional neural network algorithm would accurately diagnose hip fractures across heterogeneous populations and a web application could facilitate bedside adoption.

METHODS

Development cohort

The development cohort comprised PXR (anterior-posterior view) of patients presenting to Chang Gung Memorial Hospital's (CGMH) level I trauma center between August 2008 and December 2016. After designating unique identifiers to correlate PXRs with patient demographics and confirmatory hip fracture diagnoses (radiologist report on advanced imaging (eg, CT) or operative finding), a Python script stored anonymized PXR for imaging analysis. We excluded PXR with positioning errors (eg, full pelvis not filmed) and concurrent non-hip fractures (eg, pelvic fracture, femoral shaft fracture).

Computer vision algorithm

Our convolutional neural network structure was Xception.[9] Convolutional neural network is a deep learning methodology that permits image pattern recognition based on filters applied serially to groups of pixels. Xception optimizes image classification accuracy. The initial cohort was divided into training and internal validation images. Image augmentation methods (eg, blur, brightness, color jitter, contrast adjustment, noise addition, cropping, rotation, shifting, and zooming) were randomly applied during training to overcome possible image variations across institutions. Gradient-weighted

To cite: Choi J, Hui JZ, Spain D, *et al*. *Trauma Surg Acute Care Open* 2021;**6**:e000705.

class activation mapping overlaid heatmaps to highlight suspected fracture locations. We used TensorFlow V.1.14.0 and Keras V.2.3.1 open-source libraries on Python V.3.6.9 (Python Software Foundation).

Validation cohort

The external validation cohort comprised 500 PXRs from CGMH (n=250) and Stanford (n=250). For Stanford's cohort, 140 PXRs with confirmed hip fractures and 110 negative controls were randomly selected among injured patients presenting between January and November 2019. The CGMH validation cohort comprised same number of randomly sampled PXRs with and without confirmed hip fractures in 2017. We evaluated model performance using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, negative predictive value, positive predictive value, accuracy and F1 scores. The F1 score is a function of model precision (true positives/(true positives+false positives)) and recall (true positives/(true positives+false negatives)). Measuring precision is important when the cost of false positive is high (eg, spam email), and measuring recall is important when the cost of false negative is high (eg, sepsis screen). The F1 score is a weighted average of precision and recall.

Implementation science: web application

We integrated the final algorithm within a web application to detect hip fractures. When clinicians upload a PXR reference image (eg, screenshot, JPEG, smartphone photo), the probability of the image containing a hip fracture and a heatmap localizing the fracture location would be displayed. We used R V.3.6.3 (R Core Team, Vienna, Austria) to conduct statistical analysis.

RESULTS

Development cohort

The development cohort comprised PXRs from 4235 patients, among whom 51% (n=2089) had confirmed hip fractures (table 1). Of those with hip fractures, 53% (n=1005) had trochanteric fractures, and the remainder had femoral neck fractures. All patients were Asian. Our hip fracture detection algorithm achieved AUC of 1.00 using the training data set.

Validation cohort

Model performance on CGMH validation cohort had AUC of 0.98, with accuracy, sensitivity, and specificity at the Youden index of 94%, 92%, and 96%, respectively (table 2). Model performance on Stanford's validation cohort had AUC of 0.97, with all performance metrics, except negative predictive value (88%), above 90%. At the high sensitivity cut-off point of 95%, the specificity still achieved 85%.

Table 1 Development and validation cohort characteristics

	Development cohort		Validation cohort	
	Hip fracture (n=2089)	No hip fracture (n=2146)	CGMH (n=250)	Stanford (n=250)
Age (years), mean (SD)	48.0 (23.6)	71.4 (17.3)	54.8 (20.5)	79.1 (16.0)
Male, n (%)	1307 (62.6)	929 (43.3)	171 (68.4)	90 (36.0)
Race, n (%)				
Asian	2089 (100)	2146 (100)	250 (100)	11 (4.5)
Black				5 (1.9)
Hispanic				15 (5.8)
White				200 (79.9)
Other				19 (7.8)

CGMH, Chang Gung Memorial Hospital.

Table 2 Model performance on validation cohorts

	CGMH	Stanford
AUC	0.98	0.97
Sensitivity	0.92	0.90
Specificity	0.95	0.94
Positive predictive value	0.96	0.95
Negative predictive value	0.91	0.88
Accuracy	0.94	0.92
F1 score	0.94	0.92

AUC, area under the receiver operating curve; CGMH, Chang Gung Memorial Hospital.

Web application

Our web application (Chrome browser, account: WTC2021; code: WCTCtrauma) allows users to upload PXR in multiple image formats (ie, PNG, JPEG) from various settings (screenshot from desktop, photo from smartphone).[10] Figure 1 displays example reference and analyzed images. Of note, our algorithm also detected hip fractures on PXR with existing contralateral implants.

DISCUSSION

We refined and validated a convolutional neural network algorithm that accurately detected hip fractures on PXR across heterogeneous populations. Our algorithm had good performance metrics for validation cohorts from two institutions. A web application allowed rapid, computer vision-assisted hip fracture diagnosis to supplement clinical decision-making at the bedside.

Computer vision applications are increasingly studied to assist diagnoses of various conditions.[11–13] To our knowledge, our previous work was the first computer vision application with automatic heatmap localization to detect hip fractures on PXR.[8] Global prevalence of hip fractures and potential sex and race-specific pelvic anatomy variations warranted algorithm refinement and validation across more heterogeneous populations.[1 14 15] As the initial screening modality for a diagnosis that requires urgent management, high sensitivity is paramount for hip fracture diagnosis. The sensitivity of our algorithm was 90% and 92% for validation cohorts; in comparison, the reported sensitivity of CT for diagnosing occult hip fractures is 86%.[16]

Acquiring high-volume, high-quality data is critical for developing accurate deep learning algorithms. This is challenging for medical imaging data due to patient privacy regulations, and most researchers have only developed algorithms from their own institution.[17] Subsequently, many algorithms overfit to developmental data (ie, limited performance for non-developmental data).[18] Slight image pattern differences (eg, acquisition modality, patient positioning, presence of foreign bodies) can degrade algorithm performance for data from other institutions.[19] Our preliminary external validation shows the potential to apply algorithms to other institutional data without incurring additional algorithm training or labeling costs.

Estimates suggest it takes 17 years for health research to be translated to bedside practice.[20] After developing a high-performing algorithm, our next task was implementation science—designing a practical tool for immediate bedside implementation. Deep learning deployment into clinical workflow is critical to realize clinical utility (ie, improving patient outcomes), beyond clinical validity (ie, developing accurate diagnostic test).

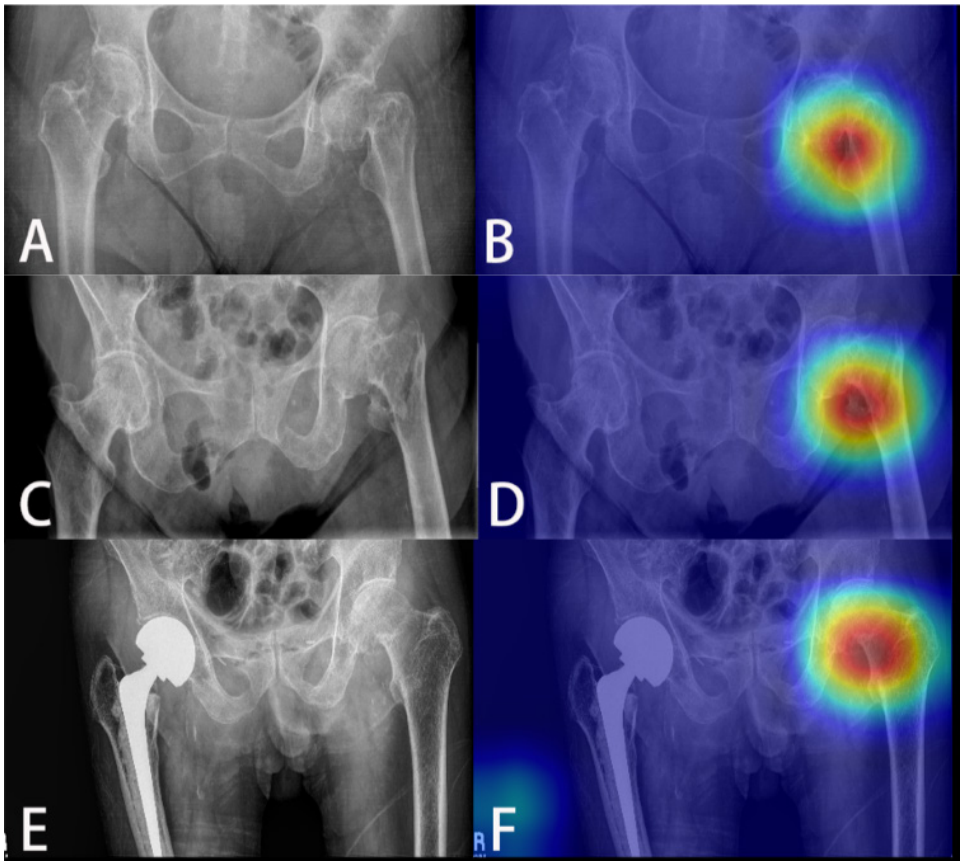

Figure 1 Detection of different types of hip fractures using computer vision algorithm. (A) Left femoral neck fracture, (B) with heatmap localization; (C) left intertrochanteric fracture, (D) with heatmap localization; (E) left femoral neck fracture with contralateral implant, (F) with heatmap localization.

Our web application allows clinicians to upload PXR from various settings with an internet connection and outputs two key results: the probability of the PXR containing a hip fracture and a heatmap highlighting the suspected fracture location for detailed review. In healthcare settings where round-the-clock radiologist interpretation is unavailable, our algorithm may be an important adjunct to rapidly detect hip fractures that require further management.

Our study has several limitations. First, our model performance, especially sensitivity, is imperfect. In spite of using our algorithm, some occult hip fractures will likely be missed. However, computer vision is meant to supplement, not replace, clinician decision-making. In select healthcare settings, benefits of using our algorithm will outweigh the minimal user costs (time to upload an image). Second, although more diverse than the development cohort, Stanford's validation cohort largely comprised Caucasian patients. This precluded validating algorithm performance for all sex-race subgroups. Whether subgroup differences in pelvic anatomy affect algorithm performance remains unclear. Third, implementing our algorithm requires thoughtful consideration of institution-specific hip fracture diagnosis pathways. For example, in an emergency department without in-house radiologists overnight, clinicians may use our algorithm to triage which PXR should be prioritized for urgent confirmatory diagnoses or further workup. However, additive benefit of using our algorithm is likely limited in settings where radiologists are readily available to interpret plain radiographs. Fourth, the algorithm is limited to detecting hip fractures and may be misleading for other concomitant PXR findings (eg, pelvic/femoral shaft/periprosthetic fractures, bone tumors). Fifth, our web application user interface is suboptimal for mobile phones. Uploading smartphone photos is feasible, but our current application is hosted on the web and does not have a mobile application counterpart. Developing a mobile application is the next step of our work. Lastly, our study only assessed an algorithm's clinical validity, not clinical utility. The ultimate goal of diagnostic tools should be to improve patient outcomes. Algorithm validation on broader target populations (eg, patients presenting with clinical concerns for hip fractures across various healthcare settings) and formal clinical utility assessment is required to evaluate whether our algorithm can improve outcomes.

CONCLUSION

We refined and validated a high-performing computer vision algorithm to detect and localize hip fractures on PXR. A web application facilitates algorithm use at the bedside, but the benefit of using our algorithm to supplement decision-making is likely institution dependent. Despite need for further validation and a more user-friendly mobile application, we are hopeful our work can exemplify strategies to incorporate implementation science and maximize computer vision's potential to improve care for injured patients.

Acknowledgements JC would like to thank the Neil and Claudia Doerhoff Fund for support of his scholarly activities.

Contributors JC, DS, and CHL contributed to study conception. JC and JZH contributed to article writing. JC, JZH, YSS, and CTC contributed to data acquisition and analysis. DS, YSS, CTC, and CHL contributed to critical review of the article.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Ethics approval Stanford University and Chang Gung Memorial Hospital Institutional Review Boards approved this study and waived need for consent.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon request

ORCID iDs
Jeff Choi http://orcid.org/0000-0003-1639-8781
Chi-Tung Cheng http://orcid.org/0000-0002-2697-4642

REFERENCES

1 Kanis JA, Odén A, McCloskey EV, Johansson H, Wahl DA, Cooper C. A systematic review of hip fracture incidence and probability of fracture worldwide. *Osteoporos Int* 2012;23:2239–56.
2 Johnell O, Kanis JA. An estimate of the worldwide prevalence and disability associated with osteoporotic fractures. *Osteoporos Int* 2006;17:1726–33.
3 Veronese N, Maggi S. Epidemiology and social costs of hip fracture. *Injury* 2018;49:1458–60.
4 Christensen L, Iqbal S, Macarios D, Badamgarav E, Harley C. Cost of fractures commonly associated with osteoporosis in a managed-care population. *J Med Econ* 2010;13:302–13.
5 Bhandari M, Swiontkowski M. Management of acute hip fracture. *N Engl J Med* 2017;377:2053–62.
6 Chellam WB. Missed subtle fractures on the trauma-meeting digital projector. *Injury* 2016;47:674–6.
7 Lampart A, Arnold I, Mäder N, Niedermeier S, Escher A, Stahl R, Trumm C, Kammerlander C, Böcker W, Nickel C, *et al*. Prevalence of fractures and diagnostic accuracy of emergency X-ray in older adults sustaining a low-energy fall: a retrospective study. *J Clin Med* 2019;9:E97:97. [Epub ahead of print: 30 Dec 2019].
8 Cheng C-T, Ho T-Y, Lee T-Y, Chang C-C, Chou C-C, Chen C-C, Chung I-F, Liao C-H. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. *Eur Radiol* 2019;29:5469–77.
9 Chollet F, 2017. Xception: deep learning with depthwise separable convolutions. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
10 PelviXNet. Hip fracture detection system. http://medcare.ym.edu.tw:8080/#/login (14 Jan 2021).
11 Kolanu N, Silverstone EJ, Ho BH, Pham H, Hansen A, Pauley E, Quirk AR, Sweeney SC, Center JR, Pocock NA. Clinical utility of computer-aided diagnosis of vertebral fractures from computed tomography images. *J Bone Miner Res* 2020;35:2307–12.
12 Le EPV, Wang Y, Huang Y, Hickman S, Gilbert FJ. Artificial intelligence in breast imaging. *Clin Radiol* 2019;74:357–66.
13 Ritchie AJ, Sanghera C, Jacobs C, Zhang W, Mayo J, Schmidt H, Gingras M, Pasian S, Stewart L, Tsai S, *et al*. Computer vision tool and technician as first reader of lung cancer screening CT scans. *J Thorac Oncol* 2016;11:709–17.
14 Handa VL, Lockhart ME, Fielding JR, Bradley CS, Brubaker L, Cundiff GW, Ye W, Richter HE. Racial differences in pelvic anatomy by magnetic resonance imaging. *Obstet Gynecol* 2008;111:914–20.
15 Lewis CL, Laudicina NM, Khuu A, Loverro KL. The human pelvis: variation in structure and function during gait. *Anat Rec* 2017;300:633–42.
16 Sadozai Z, Davies R, Warner J. The sensitivity of CT scans in diagnosing occult femoral neck fractures. *Injury* 2016;47:2769–71.
17 Kaissis GA, Makowski MR, Rückert D, Braren RF, Secure BRF. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat Mach Intell* 2020;2:305–11.
18 Rueckert D, Schnabel JA. Model-based and data-driven strategies in medical image computing. *ArXiv190910391 Cs*. 2019. http://arxiv.org/abs/1909.10391 (30 Jan 2021).
19 Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. ArXiv190912475 Cs Stat. 2019. http://arxiv.org/abs/1909.12475 (30 Jan 2021).
20 Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med* 2011;104:510–20.