European Radiology
https://doi.org/10.1007/s00330-019-06639-1

MUSCULOSKELETAL

Ruling out rotator cuff tear in shoulder radiograph series using deep learning: redefining the role of conventional radiograph

Youngjune Kim[1] · Dongjun Choi[1] · Kyong Joon Lee[1] · Yusuhn Kang[1] · Joong Mo Ahn[1] · Eugene Lee[1] · Joon Woo Lee[1] · Heung Sik Kang[1]

Received: 25 October 2019 / Revised: 25 November 2019 / Accepted: 17 December 2019

Abstract

Objective To develop a deep learning algorithm that can rule out significant rotator cuff tear based on conventional shoulder radiographs in patients suspected of rotator cuff tear.

Methods The algorithm was developed using 6793 shoulder radiograph series performed between January 2015 and June 2018, which were labeled based on ultrasound or MRI conducted within 90 days, and clinical information (age, sex, dominant side, history of trauma, degree of pain). The output was the probability of significant rotator cuff tear (supraspinatus/infraspinatus complex tear with > 50% of tendon thickness). An operating point corresponding to sensitivity of 98% was set to achieve high negative predictive value (NPV) and low negative likelihood ratio (LR−). The performance of the algorithm was tested with 1095 radiograph series performed between July and December 2018. Subgroup analysis using Fisher's exact test was performed to identify factors (clinical information, radiography vendor, advanced imaging modality) associated with negative test results and NPV.

Results Sensitivity, NPV, and LR− were 97.3%, 96.6%, and 0.06, respectively. The deep learning algorithm could rule out significant rotator cuff tear in about 30% of patients suspected of rotator cuff tear. The subgroup analysis showed that age < 60 years ($p < 0.001$), non-dominant side ($p < 0.001$), absence of trauma history ($p = 0.001$), and ultrasound examination ($p < 0.001$) were associated with negative test results. NPVs were higher in patients with age < 60 years ($p = 0.024$) and examined with ultrasound ($p < 0.001$).

Conclusion The deep learning algorithm could accurately rule out significant rotator cuff tear based on shoulder radiographs.

Key Points

- *The deep learning algorithm can rule out significant rotator cuff tear with a negative likelihood ratio of 0.06 and a negative predictive value of 96.6%.*
- *The deep learning algorithm can guide patients with significant rotator cuff tear to additional shoulder ultrasound or MRI with a sensitivity of 97.3%.*
- *The deep learning algorithm could rule out significant rotator cuff tear in about 30% of patients with clinically suspected rotator cuff tear.*

Keywords Deep learning · Rotator cuff tear · Radiography

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00330-019-06639-1) contains supplementary material, which is available to authorized users.

✉ Kyong Joon Lee
kjoon31@gmail.com

✉ Yusuhn Kang
yskang0114@gmail.com

[1] Department of Radiology, Seoul National University Bundang Hospital, 82 Gumi-ro, 173 Beon-gil, Bundang-gu, Seongnam-si, Gyeonggi-do 13620, South Korea

Abbreviations

AUC	Area under the receiver operating characteristic curve
CNN	Convolutional neural network
$Cutoff_{98\%}$	Cutoff point for an expected sensitivity of 98%
$Cutoff_{optimal}$	Optimal cutoff point determined by Youden's *J* statistic
DICOM	Digital Imaging and Communications in Medicine
FCN	Fully connected network

Published online: 05 February 2020

NPV	Negative predictive value
VAS	Visual analog scale

Introduction

Rotator cuff tear, among various causes of shoulder pain, is a condition in which surgical intervention may be required, whereas other causes, such as adhesive capsulitis, calcific tendinitis, or bursitis, are treated conservatively. Imaging plays a pivotal role in the diagnosis of rotator cuff tear, and shoulder ultrasound and MRI have comparable diagnostic performances [1, 2]. The requirement for ultrasound or MRI for rotator cuff evaluation is determined by clinicians based on history and physical examination. The major reason for performing advanced imaging in suspected rotator cuff tear is to determine the extent of rotator cuff tear and to aid orthopedic surgeons in deciding whether rotator cuff repair should be performed or not.

Shoulder radiography is routinely performed as part of the initial evaluation of patients with suspected rotator cuff tear [3, 4]. However, currently, it has a limited role in assessing rotator cuff tear and is used to exclude other possible conditions such as calcific tendinitis, osteoarthritis, or fracture. A few studies have indicated that radiographic findings, such as bone changes at the greater tubercle of the humerus and subacromial spurs, are associated with rotator cuff tear [5–8], whereas others have reported that radiographs have limited value in assessing rotator cuff tear [9–11]. When interpreted by radiologists, conventional shoulder radiographs have a wide range of diagnostic performance, with a reported negative likelihood ratio of 0.27–0.75 and a sensitivity of 62.5–91.7% [12, 13].

For conventional radiographs to be a meaningful initial screening test in the diagnosis of rotator cuff tear, it would be ideal if the negative likelihood ratio is low and the negative predictive value (NPV) is high so that rotator cuff tear could be confidently excluded without further imaging studies. In addition, the sensitivity of conventional radiographs should be reasonably high so that all patients with significant rotator cuff tears would be further examined with shoulder ultrasound or MRI. When a deep learning algorithm is employed to diagnose a disease using medical images, the algorithm generally outputs the probability of a certain disease through Softmax function. We hypothesized that conventional radiographs may be efficiently used as a screening test in the diagnosis of rotator cuff tear by appropriately adjusting the probability cutoff in a deep learning algorithm.

The purpose of this study is to develop a deep learning algorithm that can confidently rule out significant rotator cuff tear based on conventional shoulder radiographs in patients clinically suspected of rotator cuff tear.

Materials and methods

Our institutional review board approved this study and waived the requirement for informed consent considering the retrospective study design and the use of anonymized patient data.

Dataset

From January 2015 to December 2018, 8263 patients underwent both shoulder conventional radiograph series and shoulder ultrasound and/or MRI within 90 days for suspected rotator cuff tear. All patients underwent imaging under the suspicion of a rotator cuff tear by orthopedic surgeons. Among them, a few patients finally confirmed to have tumorous or infectious pathology in their shoulder but they were included in our study, unless their shoulder ultrasound or MRI was judged to be inappropriate for the evaluation of rotator cuff tear. After reviewing the entire conventional shoulder radiograph series and shoulder ultrasound and/or MRI, 7888 cases were regarded appropriate for this study (Supplementary Appendix).

A pilot study was performed to determine the size of the training set. We measured the area under the receiver operating characteristic curve (AUC) for diagnosing significant rotator cuff tear in the validation set by incrementally increasing the size of the training set by 1000 images. We found that the AUC of the deep learning algorithm plateaued when the training set contained 6000 images. In addition, we concluded that approximately 1000 radiographs are required for the test set [14] to achieve the target diagnostic performance, i.e., a sensitivity of 95%, a specificity of 50%, a negative likelihood ratio of 0.1, and an upper confidence level of 95% for a negative likelihood ratio of 0.2 [15]. Thus, we split the dataset into the training set ($n = 6113$, January 2015 to June 2018), validation set ($n = 680$, January 2015 to June 2018), and test set ($n = 1095$, July 2018 to December 2018).

Image acquisition

At our institution, the shoulder radiograph series for suspected rotator cuff tear consists of the true anteroposterior view (Grashey view), abduction anteroposterior view, 30° caudal tilt view, supraspinatus outlet view, and axillary lateral view. All conventional radiographs were taken in the standing position. The conventional radiographs were obtained using the radiography systems of various vendors, as listed in Table 1.

Shoulder ultrasound was performed by radiologists with clinical experiences of 5–10 years in musculoskeletal radiology using two ultrasound systems from a single vendor (iU22 xMATRIX and EPIQ 7, Philips Healthcare) equipped with 5–17 MHz or 5–18 MHz linear transducers. The routine shoulder ultrasound protocol was the same as that presented in Jacobson [17].

Table 1 Baseline characteristics

	Training set ($n = 6113$)	Validation set ($n = 680$)	Test set ($n = 1095$)
Age, mean ± standard deviation	59.0 ± 11.5	59.6 ± 12.2	59.5 ± 12.0
Sex			
Female	3781 (61.9)	432 (63.5)	661 (60.4)
Male	2332 (38.1)	248 (36.5)	434 (39.6)
Dominant side			
Dominant side	3031 (49.6)	351 (51.6)	461 (42.1)
Non-dominant side	2106 (34.5)	216 (31.8)	312 (28.5)
Unknown	976 (16.0)	113 (16.6)	322 (29.4)
History of trauma			
Present	1145 (18.7)	125 (18.4)	209 (19.1)
Absent	4822 (78.9)	542 (79.7)	759 (69.3)
Unknown	146 (2.4)	13 (1.9)	127 (11.6)
Degree of pain[a]			
Mild	956 (15.6)	98 (14.4)	177 (16.2)
Moderate	2871 (47.0)	328 (48.2)	572 (52.2)
Severe	1483 (24.3)	163 (24.0)	250 (22.8)
Unknown	803 (13.1)	91 (13.4)	96 (8.8)
Conventional radiograph vendor			
GE Healthcare Discovery XR656	1918 (31.4)	203 (29.9)	364 (33.2)
Philips Healthcare Digital Diagnost	2297 (37.6)	273 (40.1)	217 (19.8)
Samsung Healthcare GC85A	104 (1.7)	6 (0.9)	209 (19.1)
Siemens Healthineers Fluorospot Compact FD	1794 (29.3)	198 (29.1)	305 (27.9)
Advanced imaging modality			
Ultrasound	4425 (72.4)	499 (73.4)	741 (67.7)
MRI	1688 (27.6)	181 (26.6)	354 (32.3)
Label			
Label 0 (normal or insignificant rotator cuff tear)	4231 (69.2)	471 (69.3)	694 (63.4)
Label 1 (significant rotator cuff tear)	1882 (30.8)	209 (30.7)	401 (36.6)
High-grade partial tear	821 (13.4)	104 (15.3)	209 (19.1)
Full-thickness tear	1061 (17.4)	105 (15.4)	192 (17.5)

Data are numbers of cases and percentages in the parentheses, unless otherwise specified

[a] Degree of pain was classified into mild (VAS score < 3.5), moderate (VAS score of 3.5–7.4), and severe (VAS score ≥ 7.5) [16]

Shoulder MRI was performed using various 1.5-T and 3-T MRI systems, including our institution's 3-T MRI systems (Achieva, Ingenia, and Ingenia CX, Philips Healthcare) equipped with dedicated shoulder receiver coils (8 channels; Sense Shoulder Coil, Philips Healthcare).

All radiographs and shoulder ultrasounds were ordered by the orthopedic shoulder surgeon based on the patient's medical history and physical examination findings. Most of the shoulder MRIs were ordered by the orthopedic shoulder surgeon, but some were obtained at an outside institution prior to the patient's visit to our hospital. Ultrasound was the preferred modality when the probability of rotator cuff tear is relatively low based on the patient's medical history and physical examination, whereas MRI was preferred when the probability of rotator cuff tear was high as it enables more detailed preoperative evaluation. Further, all diagnostic ultrasonography examinations are performed by radiologists. The use of ultrasound in the outpatient clinic by non-radiologists is strictly limited to ultrasound-guided interventional procedures.

Image labeling

All Digital Imaging and Communications in Medicine (DICOM) files were downloaded from the picture archiving and communication system and used in further analyses after anonymization.

Radiographs were labeled based on the official structured reports of shoulder ultrasound and MRI. If a patient underwent shoulder ultrasound and MRI within 90 days of a radiograph, MRI reports were preferred over ultrasound reports because MRI shows higher interobserver agreement in diagnosing rotator cuff tear [18]. In daily practice, we evaluate the supraspinatus/infraspinatus tendon complex and subscapularis tendon separately, and rotator cuff abnormality is assessed as follows: normal, tendinosis, low-grade partial tear (tear with ≤ 50% tendon thickness), high-grade partial tear (tear with > 50% tendon thickness), and full-thickness tear. Only the results of the supraspinatus/infraspinatus tendon complex were used for training the deep learning algorithm. The labels were dichotomized into 0, normal or insignificant rotator cuff abnormality (normal, tendinosis, or low-grade partial tear), and 1, significant rotator cuff tear (high-grade partial tear or full-thickness tear). This categorization was based on the fact that high-grade partial tear and full-thickness tear may require surgical repair [19–22] and may therefore benefit from advanced imaging.

Clinical information

In our preliminary study, a slight increase in the AUC was observed when clinical information was added as an input variable of the deep learning algorithm (Supplementary Appendix). Thus, we utilized clinical information as an input of the algorithm. The clinical information was retrospectively collected from electronic medical records. Age, the presence of trauma history, and whether the injured arm is the dominant arm or not were recorded because they are associated with rotator cuff tear [23, 24]. Sex and the degree of pain assessed with the visual analogue scale (VAS) were also recorded. When recording the degree of pain, we recorded VAS score of the most severe pain regardless of patients' posture or movement. We generated a separate category, namely "unknown," for missing data regarding the presence of trauma history or the side of the dominant arm. In the training of the algorithm, when data regarding the degree of pain were missing in the electronic medical records, we imputed VAS scores using the propensity scores, which is generated using the data of age, sex, presence of trauma history, and side of the dominant arm in the training set.

Deep learning algorithm

The deep learning algorithm was trained with only three (true anteroposterior view, caudal 30° tilt view, and supraspinatus outlet view) out of five views in our shoulder series. These three views are known to play a role in predicting rotator cuff tear [25–28]. After loading the DICOM files of radiographs using the Pydicom library (Python Software Foundation; version 1.2.0), the images were cropped to 10×10 cm^2 patches centered on the humeral head. To augment the training dataset, the center coordinates of the patches were altered randomly and the patches were flipped in random directions. The image patches were resized to a resolution of 512×512 pixels using bilinear interpolation.

Our in-house software implementation was based on the TensorFlow library (version 1.12) on a Linux operating system (Ubuntu 16.04) with CUDA/cuDNN (versions 9.0 and 7.1, respectively) for graphic processing unit acceleration. First, three streams of convolutional neural networks (CNNs) were trained independently, one for each view of a conventional radiograph as an input image. Each CNN comprised of a stack of six squeeze-and-excitation ResNet modules [29]. The architectures of the three CNNs were identical. After training the CNNs, their output features were concatenated with the clinical information (i.e., age, sex, presence of trauma history, degree of shoulder pain, and whether the injured arm is the dominant arm or not). The concatenated features were fed to two layers of a fully connected network (FCN) to predict the probability of each label computed by 2-way Softmax function. The Xavier initialization was adopted to assign the initial network weights [30]. Cross-entropy loss was minimized by employing RMSProp optimizer [31]. L2 regularization was applied to prevent overfitting. The learning rate started from 0.01 and decayed every 5000 steps at a rate of 0.94. The minibatch size was set as 12 for the three CNNs and 20 for the FCN.

Saliency mapping was used to determine what aspects of the input images the CNN was using to decide on the presence of significant rotator cuff tear in the three views mentioned above [32, 33]. In training of each CNN, the probability of each label was calculated by using a 2-way Softmax function after Log-Sum-Exp pooling was applied on the output of the last squeeze-and-excitation ResNet module. The saliency mapping was obtained by resizing the output just before applying Log-Sum-Exp pooling to 512×512 pixels using bilinear interpolation [34]. Since the final output of each CNN was calculated by applying the 2-way Softmax function, the output just before applying Log-Sum-Exp pooling was composed of two feature maps corresponding to label 0 and label 1, respectively. These two feature maps were shown with sensitive regions for label 0 and label 1, respectively.

Statistical analysis

The AUC, sensitivity, specificity, positive predictive value, NPV, positive likelihood ratio, and negative likelihood ratio of the deep learning algorithm were measured. Two cutoff points of probability were determined in the validation set and applied to the test set, i.e., a cutoff point for an expected sensitivity of 98% (cutoff$_{98\%}$) and an optimal cutoff point determined by Youden's J statistic (cutoff$_{optimal}$). Even though the algorithm was not trained to diagnose subscapularis tear, the diagnostic performance of the algorithm with cutoff$_{98\%}$ in ruling out significant subscapularis tear (tear exceeding 50% of the tendon thickness) was also calculated. For the calculation of 95% confidence intervals for AUC, an asymptotic calculation based on DeLong et al [35] was used; for sensitivity, specificity, positive predictive value, and NPV, the Clopper-Pearson method [36] was used; and for likelihood ratios, a method proposed by Simel et al [14] was used.

Subgroup analysis was performed to identify the variables associated with negative test results (proportion of cases that the deep learning algorithm classified as label 0) and the variables associated with the NPV using Fisher's exact test. The tested variables were age, sex, presence of trauma history, degree of pain, whether current shoulder pain occurred at the dominant side or not, conventional radiograph vendor, and advanced imaging modality. The ages of patients were dichotomized into < 60 years and ≥ 60 years because age ≥ 60 years is a risk factor for rotator cuff tear [24]. The degree of pain was classified into mild (VAS score < 3.5), moderate (VAS score of 3.5–7.4), and severe (VAS score ≥ 7.5) [37]. If a certain subgroup is associated with the negative test results, we can conclude that such subgroup would be more beneficial from the deep learning algorithm. If a certain subgroup is associated with the NPV, a subgroup with higher NPV would be more beneficial from the deep learning algorithm. On the other hand, we should be cautious when the deep learning algorithm excludes the rotator cuff tear in patients in the subgroup in which the deep learning algorithm showed a significantly lower NPV.

All statistical analyses were performed using Stata 14.0 (StataCorp, College Station). Two-sided p value < 0.05 was considered to be statistically significant.

Results

The baseline characteristics and the distribution of labels in the training, validation, and test sets are summarized in Table 1.

The AUC of the deep learning algorithm in the test set was 0.91 (95% confidence interval, 0.89–0.93). The diagnostic performance of the algorithm in terms of the two cutoff points is summarized in Table 2. With cutoff$_{98\%}$, the algorithm could rule out significant rotator cuff tear in 29.9% (327/1095; 95% confidence interval, 27.2–32.7%) cases in the test set. It could confidently rule out significant rotator cuff tear with an NPV of 96.6% (316/327, 94.1–98.3%) and a negative likelihood ratio of 0.06 (0.030.11). Simultaneously, it could suggest significant rotator cuff tear with sensitivity of 97.3% (390/401, 95.1–98.6%). Especially, in patients with full-thickness rotator cuff tear, the deep learning algorithm showed a sensitivity of 100% (192/192, 98.1–100.0%). With cutoff$_{\text{optimal}}$, the algorithm showed sensitivity of 78.1% (313/401, 73.7–82.0%), specificity of 87.0% (604/694, 84.3–87.3%), positive likelihood ratio of 6.02 (4.93–7.35), and negative likelihood ratio of 0.25 (0.21–0.30).

The prevalence of significant subscapularis tear was 18.4% (201/1095, 16.1–20.8%) in our test set (Table 3). The prevalence of concurrent significant supraspinatus/infraspinatus complex tear in patients with significant subscapularis tear was 73.6% (148/201, 67.0–79.6%). The algorithm with cutoff$_{98\%}$ could also confidently rule out significant subscapularis tear with an NPV of 94.8% (310/327, 91.8–96.9%) and a negative likelihood ratio of 0.24 (0.15–0.39). It could diagnose significant subscapularis tear with a sensitivity of 91.5% (184/201, 86.8–95.0%).

The qualitative assessment of the algorithm was performed using saliency mapping [32, 33]. The most sensitive region in radiographs with label 0 (normal or insignificant rotator cuff tear) was the undersurface of the acromion, the greater tubercle of the humerus, and glenohumeral joint (Fig. 1). The most sensitive region in radiographs with label 1 (significant rotator cuff tear) was the undersurface of the acromion and the greater tubercle of the humerus (Fig. 2).

The subgroup analysis showed that patients < 60 years of age (p < 0.001), patients with suspected non-dominant side rotator cuff tear (p < 0.001), patients without trauma history (p = 0.001), or patients examined with ultrasound (p < 0.001)

Table 2 Diagnostic performance of the deep learning algorithm

	Cutoff for expected sensitivity of 98%	Optimal cutoff
Sensitivity	97.3% (390/401, 95.1–98.6%)	78.1% (313/401, 73.7–82.0%)
High-grade partial tear	94.7% (198/209, 90.8–97.3%)	65.1% (136/209, 58.2–71.5%)
Full-thickness tear	100% (192/192, 98.1–100.0%)	92.2% (177/192, 87.4–95.6%)
Specificity	45.5% (316/694, 41.8–49.3%)	87.0% (604/694, 84.3–87.3%)
Positive predictive value	50.8% (390/768, 47.2–54.4%)	77.7% (313/403, 73.3–81.6%)
Negative predictive value	96.6% (316/327, 94.1–98.3%)	87.3% (604/692, 84.6–89.7%)
Positive likelihood ratio	1.79 (1.66–1.92)	6.02 (4.93–7.35)
Negative likelihood ratio	0.06 (0.03–0.11)	0.25 (0.21–0.30)

Data are percentages and nominator/denominator and/or 95% confidence interval in the parentheses

Table 3 Diagnostic performance of the deep learning algorithm with cutoff for expected sensitivity of 98% on significant subscapularis tear

	Overall significant subscapularis tear
Prevalence	18.4% (201/1095, 16.1–20.8%)
Concurrent significant SST/IST complex tear	73.6% (148/201, 67.0–79.6%)
AUC	0.75 (0.71–0.79)
Sensitivity	91.5% (184/201, 86.8–95.0%)
Specificity	34.7% (310/894, 31.6–37.9%)
Positive predictive value	24.0% (184/768, 21.0–27.1%)
Negative predictive value	94.8% (310/327, 91.8–96.9%)
Positive likelihood ratio	1.40 (1.31–1.49)
Negative likelihood ratio	0.24 (0.15–0.39)

Data are percentages and nominator/denominator and/or 95% confidence interval in the parentheses

SST supraspinatus, *IST* infraspinatus, *AUC* area under the receiver operating characteristic curve

would benefit more from the deep learning algorithm with cutoff$_{98\%}$ (Table 4). The NPVs were significantly higher in patients with age < 60 years ($p = 0.024$) and in patients examined with ultrasound ($p < 0.001$).

Discussion

We developed a deep learning algorithm that can confidently rule out significant rotator cuff tear in shoulder radiograph series. The algorithm showed a sensitivity of 97.3%, an NPV of 96.6%, and a negative likelihood of 0.06. The deep learning algorithm could rule out significant rotator cuff tear in about 30% of patients with clinically suspected rotator cuff tear. The subgroup analysis showed that patients < 60 years of age, patients with suspected non-dominant side rotator cuff tear, patients without trauma history, or patients examined with ultrasound would benefit more from the algorithm. The NPVs were significantly higher in patients with age < 60 years and patients examined with ultrasound.

It is generally accepted that negative likelihood ratios less than 0.1 provide convincing diagnostic evidence and those less than 0.2 provides strong diagnostic evidence [15]. A negative likelihood ratio of 0.06 could be achieved by setting a cutoff point with an expected sensitivity of 98%, enabling the deep learning algorithm to effectively rule out rotator cuff tear. This number is comparable to that of Ottawa ankle rule with a reported negative likelihood ratio of 0.08, which is used to rule out clinically significant fractures and to reduce unnecessary radiographic imaging [38].

At the optimal cutoff point determined by Youden's *J* statistic, the deep learning algorithm showed high diagnostic performance with a sensitivity of 78.1% and a specificity of 87.0%. Hussain et al [12] reported that the

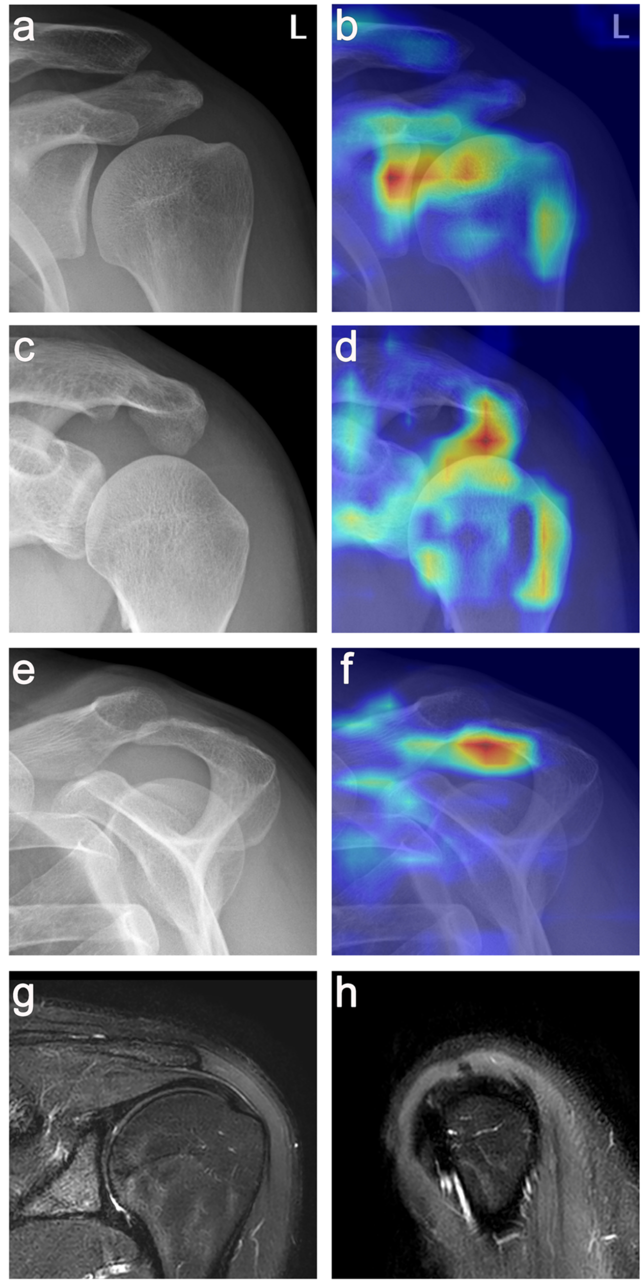

Fig. 1 Representative case for label 0 (normal or insignificant rotator cuff tear) with saliency mapping. In true anteroposterior view (Grashey view) (**a**, **b**), caudal 30° tilt view (**c**, **d**), and supraspinatus outlet view (**e**, **f**). The most highlighted regions were the undersurface of the acromion and the greater tubercle of the humerus and glenohumeral joint. MRI performed 3 weeks later revealed normal rotator cuff tendons (**g**, **h**)

sensitivity and specificity of conventional radiographs in the diagnosis of degenerative rotator cuff tear were 78.8% (95% confidence interval, 65.7–87.8%) and 77.4% (67.2–85.0%), respectively. Chin et al [13] reported a sensitivity of 62.5–91.7% and a specificity of 30.8–50.0% for conventional radiographs. Even though we did not directly compare the diagnostic performance of our deep learning algorithm with that of radiologists, the algorithm

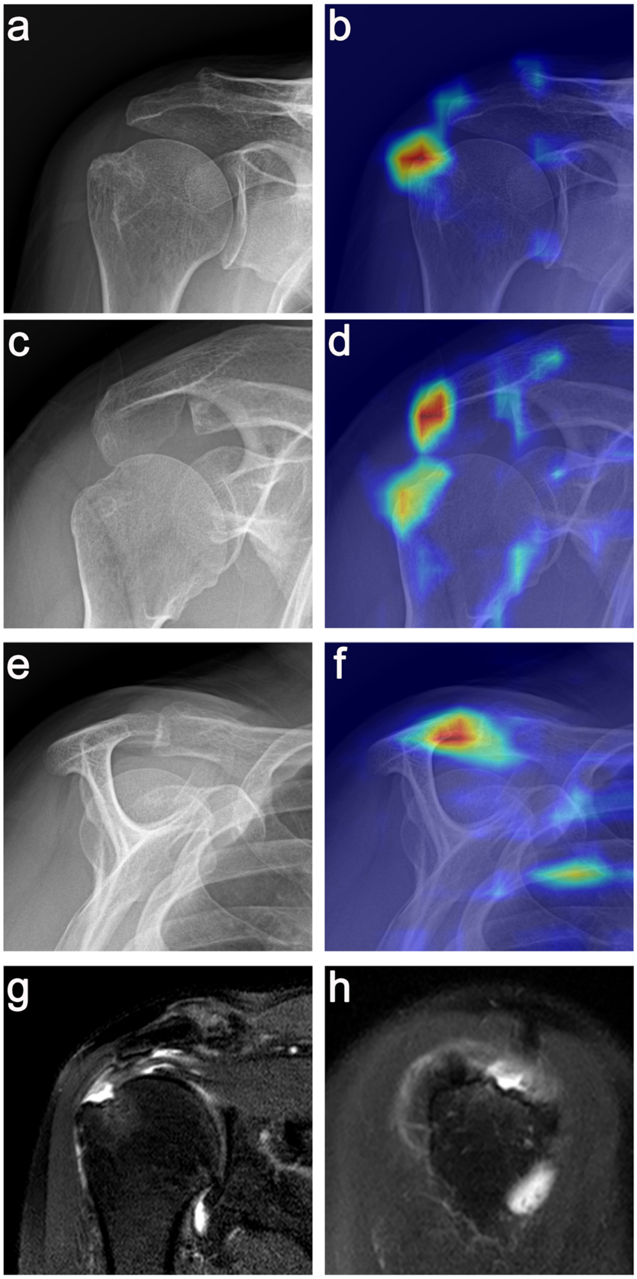

Fig. 2 Representative case for label 1 (significant rotator cuff tear) with saliency mapping. In true anteroposterior view (Grashey view) (**a**, **b**), caudal 30° tilt view (**c**, **d**), supraspinatus outlet view (**e**, **f**). The most highlighted regions were the undersurface of the acromion and the greater tubercle of the humerus. MRI performed on the same day revealed a full-thickness tear at the suprascapularis/infrascapularis conjoined tendon (**g**, **h**)

exhibited a comparable diagnostic accuracy to that of radiologists reported in the previous literature.

We did not train the deep learning algorithm regarding subscapularis tear owing to the innate low sensitivity of ultrasound and MRI in assessing the subscapularis tendon [16, 39, 40]. However, the algorithm showed a sensitivity of 91.5% in diagnosing significant subscapularis tear (tear with > 50% of tendon thickness). This is probably because 73.6% of significant subscapularis tear occurred along with supraspinatus/infraspinatus complex tear. These results were similar to the previous literature [40, 41].

The radiographic findings of rotator cuff tear include sclerosis, osteophytes, subchondral cysts, and osteolysis of the greater tubercle of the humerus [5] or bony spur at the undersurface of the acromion [6, 7]. A few researchers suggested that type III acromion may be associated with rotator cuff tear [8, 28], although this finding is controversial. Through saliency mapping, we found that the deep learning algorithm also concentrated on the changes in the greater tubercle of the humerus and the undersurface of the acromion.

The subgroup analysis showed that patients under the age of 60 years, patients with suspected non-dominant side rotator cuff tear, or patients without trauma history would benefit more from the deep learning algorithm. In particular, more than half of the patients under the age of 60 years could have avoided advanced imaging for rotator cuff evaluation. Age, history of trauma, and dominant side are known as the risk factors of rotator cuff tear [23, 24]. Hence, our deep learning algorithm was more beneficial to patients without the risk factors of rotator cuff tear, making it possible to safely rule out significant rotator cuff tear.

The NPVs of our deep learning algorithm were significantly higher in patients with age < 60 years. The diagnostic performance of a test is subject to different sources of variations such as demographic features, disease severity, and disease prevalence in the population [42]. In the statistical point of view, the NPV decreases as disease prevalence increases. Thus, a higher NPV in the subgroup with age < 60 years can be explained by the fact that the prevalence of rotator cuff tear is lower in patients with age < 60 years. Otherwise, the NPV was not associated with other clinical variables. Such results indicate that the deep learning algorithm can be applied to patients in various clinical settings, regardless of the clinical likelihood of rotator cuff tear. Furthermore, the NPV was not affected by the vendor of the conventional radiograph; this showed the robustness of the algorithm.

The association between the advanced imaging modality used as the reference standard and the negative test results and NPV is attributable to referral bias and differential verification bias. As our institution is a tertiary urban teaching hospital, patients with rotator cuff tear that may require surgical repair visit our hospital for surgery after MRI examination at a local clinic. Additionally, orthopedic surgeons select imaging modality according to the probability of rotator cuff tear based on patients' clinical information and the results of physical examination. Ultrasound is favored in patients with low probability of rotator cuff tear, whereas MRI is preferred in patients with high probability of rotator cuff tear because MRI allows for more detailed preoperative evaluation.

Table 4 Subgroup analysis to identify factors associated with negative test result and negative predictive value

	Negative test result		Negative predictive value	
	Data	*p* value	Data	*p* value
Age[a]				
< 60 years	48.6% (250/514, 44.2–53.1%)	< 0.001*	98.0% (245/250, 95.4–99.3%)	0.024*
≥ 60 years	13.3% (77/581, 10.6–16.3%)		92.2% (71/77, 83.8–97.1%)	
Sex				
Female	30.9% (204/661, 27.4–34.5%)	0.38	96.1% (196/204, 92.4–98.3%)	0.55
Male	28.3% (123/434, 24.1–32.8%)		97.6% (120/123, 93.0–99.5%)	
Dominancy				
Dominant side	18.2% (84/461, 14.8–22.1%)	< 0.001*	95.2% (80/84, 88.3–98.7%)	> 0.99
Non-dominant side	33.0% (103/312, 27.8–38.5%)		96.1% (99/103, 90.4–98.9%)	
Unknown	43.5% (140/322, 38.0–49.1%)		97.9% (137/140, 93.9–99.6%)	
Trauma history				
Present	18.2% (38/209, 13.2–24.1%)	0.001*	97.4% (37/38, 86.2–99.9%)	> 0.99
Absent	29.8% (226/759, 26.5–33.2%)		96.5% (218/226, 93.1–98.5%)	
Unknown	49.6% (63/127, 40.6–58.6%)		96.8% (61/63, 89.0–99.6%)	
Degree of pain[b]				
Mild	25.4% (45/177, 19.2–32.5%)	0.32	97.8% (44/45, 88.2–99.9%)	0.088
Moderate	30.9% (177/572, 27.2–34.9%)		98.3% (174/177, 95.1–99.6%)	
Severe	27.6% (69/250, 22.2–33.6%)		92.8% (64/69, 83.9–97.6%)	
Unknown	37.5% (36/96, 27.8–48.0%)		94.4% (34/36, 81.3–99.3%)	
Conventional radiograph vendor				
GE Healthcare Discovery XR656	26.6% (97/364, 22.2–31.5%)	0.145	96.9% (94/97, 91.2–99.4%)	0.95
Philips Healthcare Digital Diagnost	28.6% (62/217, 22.7–35.1%)		95.2% (59/62, 86.5–99.0%)	
Samsung Healthcare GC85A	29.7% (62/209, 23.6–36.4%)		96.8% (60/62, 88.8–99.6%)	
Siemens Healthineers Fluorospot Compact FD	34.8% (106/305, 29.4–40.4%)		97.2% (103/106, 92.0–99.4%)	
Advanced imaging modality				
Ultrasound	37.1% (275/741, 33.6–40.7%)	< 0.001*	98.9% (272/275, 96.8–99.8%)	< 0.001*
MRI	14.7% (52/354, 11.2–18.8%)		84.6% (44/52, 71.9–93.1%)	
Total	29.9% (327/1095, 27.2–32.7%)		96.6% (316/327, 94.1–98.3%)	

Data are percentages and nominator/denominator with 95% confidence interval in the parentheses. Missing data regarding dominant side, trauma history, and the degree of pain were excluded during the calculation of *p* values

[a] Age were dichotomized into < 60 years and ≥ 60 years, since age ≥ 60 years is a risk factor of rotator cuff tear [26]

[b] Degree of pain was classified into mild (VAS score < 3.5), moderate (VAS score of 3.5–7.4), and severe (VAS score ≥ 7.5) [16]

*p value < 0.05

Our study has several limitations. First, it is a retrospective study conducted in a single institution; this may limit the generalizability of the results. We performed subgroup analysis to find any possible association between the clinical factors and the negative test rate and NPV. Through this analysis, we partly shown the transferability of the results of the deep learning algorithm. However, further validation of the algorithm is required. Second, we did not directly train the algorithm regarding subscapularis tear. However, the algorithm could rule out significant subscapularis tear detected in ultrasound and MRI. Third, although we defined significant rotator cuff tear to include high-grade partial tear, the clinical significance of "significant rotator cuff tears" may be questionable because treatment for high-grade partial tears is not yet uniform and is usually decided based on clinical judgment. Finally, most of our labeling procedure was based on the ultrasound result, which may have substantial interobserver variability. However, several articles have shown that shoulder ultrasound is comparable to MRI in terms of diagnosis of rotator cuff tear [1, 2, 43].

In conclusion, our deep learning algorithm could accurately rule out significant rotator cuff tear using shoulder radiography. The implementation of the algorithm in the shoulder radiograph series could successfully redefine the role of conventional radiographs in the diagnosis of rotator cuff tear.

Acknowledgments The authors sincerely thank Jeongmin Choi for her contribution in the data collection.

Funding information This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2019R1F1A1060126), National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. 2017R1D1A1B03033610), and grant from the SNUBH Research Fund (No. 13-2019-006).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Yusuhn Kang.

Conflict of interest The authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors (Dongjun Choi) has significant statistical expertise.

Informed consent Written informed consent was waived by the institutional review board.

Ethical approval Institutional review board approval was obtained.

Methodology
• retrospective
• experimental
• performed at one institution

References

1. de Jesus JO, Parker L, Frangos AJ, Nazarian LN (2009) Accuracy of MRI, MR arthrography, and ultrasound in the diagnosis of rotator cuff tears: a meta-analysis. AJR Am J Roentgenol 192:1701–1707
2. Roy JS, Braen C, Leblond J et al (2015) Diagnostic accuracy of ultrasonography, MRI and MR arthrography in the characterisation of rotator cuff disorders: a systematic review and meta-analysis. Br J Sports Med 49:1316–1328
3. Amini B, Beckmann NM, Beaman FD et al (2018) ACR Appropriateness Criteria® shoulder pain-traumatic. J Am Coll Radiol 15:S171–S188
4. Small KM, Adler RS, Shah SH et al (2018) ACR Appropriateness Criteria® shoulder pain—atraumatic. J Am Coll Radiol 15:S388–S402
5. Pearsall AW 4th, Bonsell S, Heitman RJ, Helms CA, Osbahr D, Speer KP (2003) Radiographic findings associated with symptomatic rotator cuff tears. J Shoulder Elbow Surg 12:122–127
6. Hardy DC, Vogler JB 3rd, White RH (1986) The shoulder impingement syndrome: prevalence of radiographic findings and correlation with response to therapy. AJR Am J Roentgenol 147:557–561
7. Cone RO 3rd, Resnick D, Danzig L (1984) Shoulder impingement syndrome: radiographic evaluation. Radiology 150:29–33
8. Peh WC, Farmer TH, Totty WG (1995) Acromial arch shape: assessment with MR imaging. Radiology 195:501–505
9. Huang LF, Rubin DA, Britton CA (1999) Greater tuberosity changes as revealed by radiography: lack of clinical usefulness in patients with rotator cuff disease. AJR Am J Roentgenol 172:1381–1388
10. Liotard JP, Cochard P, Walch G (1998) Critical analysis of the supraspinatus outlet view: rationale for a standard scapular Y-view. J Shoulder Elbow Surg 7:134–139
11. Hyvonen P, Paivansalo M, Lehtiniemi H, Leppilahti J, Jalovaara P (2001) Supraspinatus outlet view in the diagnosis of stages II and III impingement syndrome. Acta Radiol 42:441–446
12. Hussain A, Muzzammil M, Butt F, Valsamis EM, Dwyer AJ (2018) Effectiveness of plain shoulder radiograph in detecting degenerate rotator cuff tears. J Ayub Med Coll Abbottabad 30:8–11
13. Chin K, Chowdhury A, Leivadiotou D, Marmery H, Ahrens PM (2017) The accuracy of plain radiographs in diagnosing degenerate rotator cuff disease. Shoulder Elbow. https://doi.org/10.1177/1758573217743942
14. Simel DL, Samsa GP, Matchar DB (1991) Likelihood ratios with confidence: sample size estimation for diagnostic test studies. J Clin Epidemiol 44:763–770
15. Jaeschke R, Guyatt GH, Sackett DL (1994) Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 271:703–707
16. Narasimhan R, Shamse K, Nash C, Dhingra D, Kennedy S (2016) Prevalence of subscapularis tears and accuracy of shoulder ultrasound in pre-operative diagnosis. Int Orthop 40:975–979
17. Jacobson JA (2011) Shoulder US: anatomy, technique, and scanning pitfalls. Radiology 260:6–16
18. Okoroha KR, Mehran N, Duncan J et al (2017) Characterization of rotator cuff tears: ultrasound versus magnetic resonance imaging. Orthopedics 40:e124–e130
19. Ide J, Maeda S, Takagi K (2005) Arthroscopic transtendon repair of partial-thickness articular-side tears of the rotator cuff: anatomical and clinical study. Am J Sports Med 33:1672–1679
20. Moosmayer S, Lund G, Seljom US et al (2019) At a 10-year follow-up, tendon repair is superior to physiotherapy in the treatment of small and medium-sized rotator cuff tears. J Bone Joint Surg Am 101:1050–1060
21. McCallister WV, Parsons IM, Titelman RM, Matsen FA 3rd (2005) Open rotator cuff repair without acromioplasty. J Bone Joint Surg Am 87:1278–1283
22. Harryman DT 2nd, Mack LA, Wang KY, Jackins SE, Richardson ML, Matsen FA 3rd (1991) Repairs of the rotator cuff. Correlation of functional results with integrity of the cuff. J Bone Joint Surg Am 73:982–989
23. Yamamoto A, Takagishi K, Osawa T et al (2010) Prevalence and risk factors of a rotator cuff tear in the general population. J Shoulder Elbow Surg 19:116–120
24. Sayampanathan AA, Andrew TH (2017) Systematic review on risk factors of rotator cuff tears. J Orthop Surg (Hong Kong) 25:2309499016684318
25. Koh KH, Han KY, Yoon YC, Lee SW, Yoo JC (2013) True anteroposterior (Grashey) view as a screening radiograph for further imaging study in rotator cuff tear. J Shoulder Elbow Surg 22:901–907
26. Ono K, Yamamuro T, Rockwood CA Jr (1992) Use of a thirty-degree caudal tilt radiograph in the shoulder impingement syndrome. J Shoulder Elbow Surg 1:246–252
27. Mayerhoefer ME, Breitenseher MJ, Roposch A, Treitl C, Wurnig C (2005) Comparison of MRI and conventional radiography for assessment of acromial shape. AJR Am J Roentgenol 184:671–675
28. Bigliani LU, Ticker JB, Flatow EL, Soslowsky LJ, Mow VC (1991) The relationship of acromial architecture to rotator cuff disease. Clin Sports Med 10:823–838
29. Hu J, Shen L, Albanie S, Sun G, Wu E (2018) Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7132-7141

30. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 249-256
31. Hinton G, Srivastava N, Swersky K (2012) Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Available via https://www.cs.toronto.edu/~hinton/coursera/lecture6/lec6.pdf
32. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: visualising image classification models and saliency maps. Available via https://arxiv.org/pdf/1312.6034.pdf. Accessed May 30 2019
33. Philbrick KA, Yoshida K, Inoue D et al (2018) What does deep learning see? Insights from a classifier trained to predict contrast enhancement phase from CT images. AJR Am J Roentgenol 211:1184–1193
34. Yao L, Prosky J, Poblenz E, Covington B, Lyman K (2018) Weakly supervised medical diagnosis and localization from multiple resolutions. arXiv preprint arXiv:180307703
35. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 44:837–845
36. Clopper CJ, Pearson ES (1934) The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26:404–413
37. Boonstra AM, Schiphorst Preuper HR, Balk GA, Stewart RE (2014) Cut-off points for mild, moderate, and severe pain on the visual analogue scale for pain in patients with chronic musculoskeletal pain. Pain 155:2545–2550
38. Bachmann LM, Kolb E, Koller MT, Steurer J, ter Riet G (2003) Accuracy of Ottawa ankle rules to exclude fractures of the ankle and mid-foot: systematic review. BMJ 326:417
39. Foad A, Wijdicks CA (2012) The accuracy of magnetic resonance imaging and magnetic resonance arthrogram versus arthroscopy in the diagnosis of subscapularis tendon injury. Arthroscopy 28:636–641
40. Garavaglia G, Ufenast H, Taverna E (2011) The frequency of subscapularis tears in arthroscopic rotator cuff repairs: a retrospective study comparing magnetic resonance imaging and arthroscopic findings. Int J Shoulder Surg 5:90–94
41. Ono Y, Sakai T, Carroll MJ, Lo IK (2017) Tears of the subscapularis tendon: a critical analysis review. JBJS Rev 5:1–12
42. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J (2004) Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med 140:189–202
43. Lenza M, Buchbinder R, Takwoingi Y, Johnston RV, Hanchard NC, Faloppa F (2013) Magnetic resonance imaging, magnetic resonance arthrography and ultrasonography for assessing rotator cuff tears in people with shoulder pain for whom surgery is being considered. Cochrane Database Syst Rev. https://doi.org/10.1002/14651858.CD009020.pub2:CD009020