centre hospitalier de Valenciennes

Research results - 2020/08/25
Pôle Imagerie Médicale
Dr Parsy et al

MSK – AI - RETROSPECTIVE STUDY

INTRODUCTION

The objective of the study is to evaluate the performance of Arterys' MSK AI module for the detection of abnormalities emergency department x-rays of the Valenciennes Hospital Center.

OBJECTIVES

The main objective is to :
- Verify the clinical validity of the products for musculo skeletal indications (MSK)

The secondary objectives are :
- Evaluate integration in the workflow
- Evaluate diagnostic gaps between practitioners
- Verify model performance for specific anomalies and anatomical subgroups

MATERIAL AND METHOD

The analysis of the images will be done without then with Artificial Intelligence (AI) in an isolated office with all the necessary tools for an unlimited session and without a cell phone.
The images was randomly selected to avoid image quality bias.
No time recording to avoid playback bias

Retrospective study

Selection of a retrospective population of 650 patients with a known diagnosis randomly selected. Constitution of 12 sets of images corresponding to the desired anatomical subgroups. The quality of the images will therefore also be random. Each radiologist will examine a different set of images during the readings without and with AI. (Reading A and Reading B)

Each anatomical subgroup will contain between 30 and 70% pathological cases for an overall distribution of 50 - 50% pathological and non-pathological cases. Each set of images will include, if possible, a significant proportion of the different pathologies to be detected.

The Gold standard of the study is the initial radiology report within clinical context is mentioned. The study readings will be made without clinical context (except for the traumatic context). In case of discrepancy between the initial reading and the A-reading, a agreed diagnosis will be established by 2 expert radiologists.

MSK

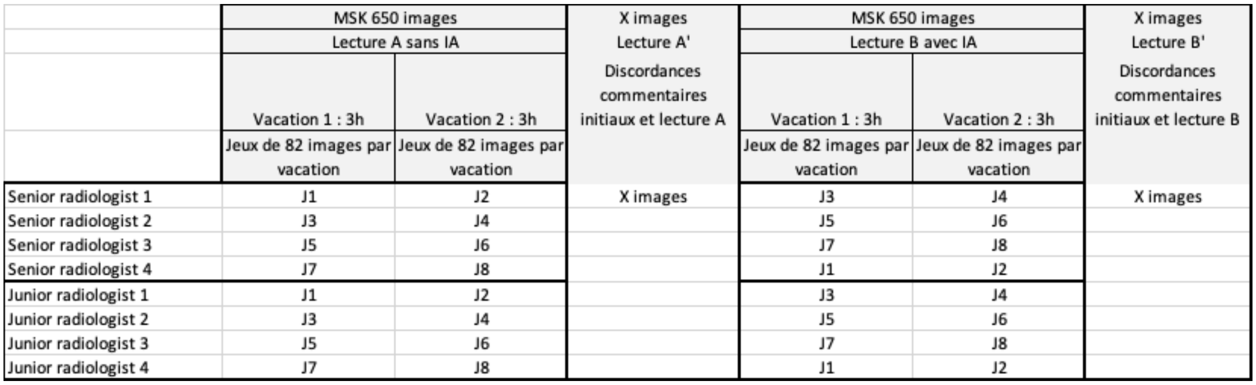
4 senior and 4 junior radiologists will read the images and analyze the differences between their diagnosis and AI

Reading cycles :

- Reading 01: Data set validation
- Reading A: Reading images with current conditions (with PACS)
- Reading A': Analysis of discrepancies between the Reading A report and the initial cases for consensual diagnosis
- Reading B: Reading images with AI (with Arterys)

- Reading B': Analysis of discrepancies in Reading B reports with initial cases for consensus diagnosis

Distribution of the 650 images of the 12 MSK anatomical data sets into 8 groups of 81/82 images per radiologist.



	MSK 650 images		X images	MSK 650 images		X images
	Lecture A sans IA		Lecture A'	Lecture B avec IA		Lecture B'
	Vacation 1 : 3h	Vacation 2 : 3h	Discordances commentaires initiaux et lecture A	Vacation 1 : 3h	Vacation 2 : 3h	Discordances commentaires initiaux et lecture B
	Jeux de 82 images par vacation	Jeux de 82 images par vacation		Jeux de 82 images par vacation	Jeux de 82 images par vacation	
Senior radiologist 1	J1	J2	X images	J3	J4	X images
Senior radiologist 2	J3	J4		J5	J6	
Senior radiologist 3	J5	J6		J7	J8	
Senior radiologist 4	J7	J8		J1	J2	
Junior radiologist 1	J1	J2		J3	J4	
Junior radiologist 2	J3	J4		J5	J6	
Junior radiologist 3	J5	J6		J7	J8	
Junior radiologist 4	J7	J8		J1	J2	

Process:

Prior validation of the relevance of each dataset by 1 radiologist (MM) not participating in the readings.

Entry of the results of each reading in the same Excel file without and with IA.

A period of one month between readings with and without AI is not necessary because the data sets will be swapped between the readers during the different readings.

A consolidation phase of the results following the 3 readings will be necessary for the analysis of the results of the junior and senior radiologists without and with AI for the 2 groups.

Cases will be divided into the following anatomical groups and subgroups: Lower limbs (Pelvis/Bassin, Ankle/Cheville, Knee/Genou, Hip/Hanche, Leg/Jambe, Foot/Pied), Upper limbs (Arm/Bras, Elbow/Coude, Shoulder/Epaule, Hand/Main, Wrist/Poignet), Thorax (Ribs/Côtes)

Composition of data set :

Historical data extracted from PACS (June to September 2019)
Proposed Study Data Sets and Samples

The proportion of cases within each subgroup is proportional to the total number of cases, and the distribution into upper and lower limbs is also proportional. The number of images is sufficient to constitute a representative sample of the original population**.**

Statistical Analysis :

Variables were described in terms of frequencies and percentages. Age was expressed in terms of mean and standard deviation.
The sensitivity and specificity of Artificial Intelligence, Junior Radiologists and Senior Radiologists to the Gold Standard were calculated with their 95% confidence interval. The Mc Nemar test was used for the comparison of sensitivities, as well as for the comparison of specificities between 2 radiologists.
The level of significance was set at 5%. Statistical analyses were performed using SAS software (SAS Institute version 9.4).

RESULTS :

Gold Standard (GS)
Population : 620
Age :
avg 53.6 +-23.9 (18-98)
med 51.0 (32.0; 75.0)

Fracture = 253 out of 620 (40.8%)
Dislocation = 28 out of 620 (4.5%)

Effusion = 25 out of 69 (36.2%) (only analyzable for elbow, knee and ankle)

Fracture
IA versus GS: Contingency table

	Overall	1	2	3	4	5	6	7	8	9	10	11	12
Body Part		Elbow	Knee	Foot	Hand	Hip	Leg	Pelvis	Ankle	Arm	Ribs	Shoulder	Wrist
total	620	30	48	71	90	67	33	65	73	30	9	60	44
Patho	253	4	7	29	51	27	6	31	26	10	1	33	28
Spe	90	84	87	92	92	92	96	82	93	100	100	81	93
Sen	77	100	57	51	70	81	100	74	84	90	100	84	96
NPV	85	100	92	73	70	88	100	77	91	95	100	81	93
PPV	85	50	44	83	92	88	85	79	88	100	100	84	96
FP	9	15	42	48	29	18	0	25	15	10	0	15	3
FN	22	0	12	7	7	7	3	17	6	0	0	18	6

Significance threshold at 5%.

Contingency Table Radiologists +- iA vs GS

	Sénior	**Sénior +iA**	**Junior end study**	**Junior +iA**
Overall				
Effectif	544	310	546	310
Pato	214	108	230	145
Spe	98	99	97	98
Sen	93	95	92	95
NPV	96	97	94	96
PPV	97	99	96	97
FP				
FN				

Comparison of sensitivity/specificity (McNemar's test) :

JR significantly more sensitive than iA, p value <0.0001

JR significantly more specific than iA, p value =0.0002

SR significantly more sensitive than iA, p value < 0.0001

SR significantly more specific than iA, p value <0.0001

SR no more sensitive than JR, p value =0.1336

SR no more specific than JR, p value =0.5637

Comparison not applicable for : SR+iA vs SR, JR+iA vs JR, SR vs JR+iA.

DISCUSSION

An artificial intelligence algorithm to detect fractures on X-rays from ED department would be interesting to enable faster and more appropriate patient management. Bone X-rays in emergency departments represent a challenge in France because the majority of hospital are not able to provide an interpretation by a radiologist, which is a legal obligation. As a result, only emergency physicians interpret x-rays without a second reading by a radiologist, which sometimes leads to missed diagnoses and inappropriate treatment. Our organisation is one of the rare hospitals where x-rays are read by radiologist (24h delayed) Discordant diagnoses, which can be empirically estimated at 5%, are then reported to the emergency department so that the treatment can be adapted.

Fracture search is the overwhelming indication in bone X-rays from the emergency department.

In the literature, only few studies have focused on AI in emergency department bone radiographs (Beyaz and al. 2020, Kim and al. 2018). This is due to different practices and needs in different countries. For example, in the USA, when fractures are suspected, the use of CT scans is more frequent.
On the other hand, there are several studies on chest and dental X-rays.

In the study by Beyaz and al., which focuses only on femoral neck fractures, the sensitivity is 83% slightly higher than our algorithm, but the specificity is lower, evaluated at 73%, compared to 90% in our study. These sensitivity and specificity values are complementary and are adjusted by the detection threshold applied to the algorithm.Indeed, the algorithm can be asked to be more sensitive, but then there will be more false negatives, and the specificity will be higher.

Indeed, the algorithm can be asked to be more sensitive, but then there will be more false negatives, and the specificity will decline.

Most of the studies carried out are highly targeted, and focus only on one anatomical zone (femoral neck: Beyaz and al. , Chung and al. humerus, Kim and al. wrist). Our study is the only one currently concerning all parts of the peripheral skeleton.

In our study, the algorithm is compared with senior (osteoarticular experts) and junior (final year in osteoarticular position) radiologists. This explains the high performance of the junior radiologists compared to the senior radiologists, and in comparison to the algorithm. Comparison with non-radiologists and interns at the beginning of the course could reveal a lower level of performance.

Moreover, the conditions of the study: in specific shift, and knowing that the results are going to be analyzed and compared to an AI algorithm may bias the results, increasing the performance of juniors, particularly motivated and diligent in the detection of anomalies.

The study by anatomical region provides interesting information. The low prevalence of fractures for the anatomical regions "elbow / knee / ribs" does not allow the results to be interpreted. Indeed, out of the 9 cases of rib radiographs, only 1 fracture was present, so the sensitivity and specificity of 100% is not exploitable. Conversely, the diagnostic performance of the algorithm for the upper limb, for example for the arm (sensitivity of 90% and specificity of 100%) and for the wrist (sensitivity of 96% and specificity of 93%), for a prevalence of pathological cases of 33 and 67% seem particularly encouraging. Thus, this information by anatomical region also makes it possible to target areas of AI performance and gap areas in order to adapt the algorithm and train it on specific areas.The relatively small number of 650 files makes it impossible to exploit. We will therefore have to confirm our results with a study of several thousand cases.

Furthermore, the algorithm is not adapted to the detection of spinal fractures, but this indication is increasingly benefiting from a CT scan from the outset.

CONCLUSION

Currently, the MSK milvue software has a lower performance than a senior and junior radiologist for the detection of abnormalities (fractures, dislocation, effusion) on standard x-rays from the emergency department. The diagnostic performance of the senior and junior radiologist is improved with the help of AI, but not significantly.

Nevertheless, these performances are interesting and allow to handle with confidence a prospective study on the field.
ED physicians are in the front line for reading X-rays and do not have the background of radiologists. Thus, it will be interesting to evaluate the contribution of the AI software to emergency physicians in daily practice.

REFERENCES

- Khan SH, Hedges WP. Workload of consultant radiologists in a large DGH and how it compares to international benchmarks. Clin Radiol. 2013 May;68(5):e239-44. doi: 10.1016/j.crad.2012.10.016.
- von Schacky CE, Sohn JH, Liu F, Ozhinsky E, Jungmann PM, Nardo L, Posadzy M, Foreman SC, Nevitt MC, Link TM, Pedoia V. Development and Validation of a Multitask Deep Learning Model for Severity Grading of Hip Osteoarthritis Features on Radiographs. Radiology. 2020 Apr;295(1):136-145. doi: 10.1148/radiol.2020190925.
- Beyaz S, Açıcı K, Sümer E. Femoral neck fracture detection in X-ray images using deep learning and genetic algorithm approaches. Jt Dis Relat Surg. 2020;31(2):175-183. doi: 10.5606/ehc.2020.72163.
- Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, Ibrahim A, Hulst LV, Lambin P, Poeze M. Deep learning in fracture detection: a narrative review. Acta Orthop. 2020 Apr;91(2):215-220. doi: 10.1080/17453674.2019.1711323.
- Kim D H, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol 2018; 73: 439-45.
- Chung S W, Han S S, Lee J W, Oh K S, Kim N R, Yoon J P, Kim J Y, Moon S H, Kwon J, Lee H J, Noh Y M, Kim Y. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop 2018; 89: 468-73.