TABLE DES MATIERES

TABLE DES TABLEAUX

TABLE DES FIGURES

INTRODUCTION

Les suspicions de fractures du squelette appendiculaire représentent le premier motif de consultation au sein des urgences pédiatriques (1), avec un taux correspondant à plus d'un tiers des admissions. L'exploration de ces patients repose sur l'imagerie médicale et plus particulièrement, en première intention, sur la radiographie standard (2).
Cette modalité d'imagerie est relativement accessible tant sur le plan économique que technique. Elle représente, de ce fait, la principale méthode d'imagerie dans le monde (3)(4).

Bien qu'accessible, l'interprétation des radiographies traumatiques du squelette appendiculaire demande une expertise certaine, a fortiori lorsqu'il s'agit de radiographies pédiatriques où les différentes phases de croissance osseuse et les variantes anatomiques peuvent prêter à confusion et se révéler être une source d'erreurs diagnostiques (5)(6).
Le territoire français présente une carence en médecins spécialisés en radiologie avec une forte disparité territoriale. Cette disparité est encore plus marquée quand il s'agit de radiologie-pédiatrique, discipline principalement pratiquée dans les centres hospitalo-universitaires (7)(8).

Ce manque de praticiens spécialisés se traduit fréquemment par l'interprétation des radiographies pédiatriques ostéo-articulaires, dans un contexte d'urgence, par les médecins urgentistes sans avis immédiat ou secondaire radio-pédiatrique ou à minima radiologique (2). Ce manque d'expertise radiologique est source d'erreurs de diagnostic (9)(10).

Ces dernières représentent jusqu'à 80 % (11) des erreurs commises dans les services d'accueil des urgences et sont à l'origine de morbidités supplémentaires (12). Une mauvaise interprétation d'une radiographie peut, en effet, entrainer de graves complications tel qu'un cal vicieux, diminuant l'amplitude des mouvements du patient, une arthrose post-traumatique ou encore un effondrement articulaire, pouvant nécessiter à terme une chirurgie (2).

Outre une morbidité augmentée, les erreurs de diagnostic relatives aux fractures du squelette appendiculaire représentent une cause fréquente de dépôt de plainte contre les établissements hospitaliers et les professionnels de santé. Aux États-Unis d'Amérique, ces erreurs de

diagnostic représentent le deuxième motif d'action en justice, après les néoplasies mammaires (13).

Apporter un dispositif d'aide au diagnostic aux urgentistes et permettre à un plus grand nombre de radiologues d'être assistés pour l'interprétation des radiographies pédiatriques traumatiques, permettrait d'éviter certains incidents et poursuites judiciaires.

L'essor que connaît, depuis de nombreuses années, la conception des logiciels d'Intelligence Artificielle (IA) aidant à la détection d'anomalies sur les examens d'imagerie médicale constitue, dans ce contexte, une piste à explorer. Ces logiciels ont notamment été développés pour la détection des néoplasies mammaires sur mammographies (14) et des nodules pulmonaires sur examens tomodensitométriques (15).

De surcroît, l'avènement récent du *deep-learning* permet d'atteindre des performances diagnostiques intéressantes concernant plusieurs modalités d'imagerie et de détection de pathologies diverses (16)(17). Des études ont montré l'intérêt de l'utilisation de logiciels d'IA, d'aide au diagnostic, dans l'interprétation de radiographies pour suspicion de fracture du squelette appendiculaire chez l'adulte (7) ou, de certaines articulations ciblées, chez l'enfant (18)(19)(20).

À l'heure actuelle, en radiopédiatrie, seuls des logiciels permettant une évaluation automatique et systématisée de l'âge osseux sont utilisés en pratique courante (21). Concernant la détection de fracture des membres chez l'enfant, quelques études isolées existent (18)(20). Bien que ciblées principalement sur l'articulation du coude et notamment les fractures supra-condyliennes, elles affichent des performances diagnostiques remarquables avec une sensibilité de détection oscillant entre 91% et 93% et une spécificité allant de 84 à 91%.

La présente étude analyse les performances et la concordance diagnostique relatives aux examens radiographiques recherchant une fracture du squelette appendiculaire, avec comme *gold standard* l'interprétation d'un senior de radiopédiatrie ayant plus de 20 ans d'expérience en comparaison avec un Interne de radiologie en fin de cursus, pouvant être assimilé à un radiologue généraliste, avec et sans le concours d'un logiciel d'IA d'aide à la détection de fracture.

Les données de performance et concordance diagnostique ont pu être évaluées via l'interprétation de quatre séries de 100 examens radiologiques pédiatriques traumatiques pour les principales articulations que sont le genou, la cheville, le coude et le poignet, issues de patients âgés de moins de 16 ans, reçus au Service d'Accueil des Urgences Pédiatriques du Centre Hospitalier Universitaire de l'hôpital Nord de Marseille. Ces radiographies ont été réalisées en 2020.

L'objectif de cette étude comparative est d'évaluer l'apport de l'IA seule ou en soutien d'un radiologue non spécialisé en radiopédiatrie dans le cadre de lésions traumatiques du squelette appendiculaire de l'enfant.

MATERIELS ET METHODES

CONFLITS D'INTERETS ET AFFILIATION A L'ENTREPRISE GLEAMER

L'étude en objet, a été menée et écrite par des praticiens hospitaliers non rémunérés et non affiliés à l'entreprise Gleamer. Ladite entreprise, développant le logiciel d'IA, n'a financé par aucun moyen les différentes étapes de l'étude. En outre, les professionnels de santé n'ont aucun conflit d'intérêts à déclarer en lien avec le sujet d'étude.

AUTORISATION DE L'ETUDE ET PROTECTION DES DONNEES PERSONNELLES

La présente étude a fait l'objet d'une approbation préalable par le Conseil d'Administration de l'Hôpital Nord de Marseille. Conformément au RGPD[1], les modalités de mise en place de l'étude n'ont pas nécessité le consentement du patient. En conséquence, un numéro RGPD a été attribué à l'étude par le délégué à la protection des données de l'A.P.H.M.

LOGICIEL D'INTELLIGENCE ARTIFICIELLE D'AIDE A LA DETECTION DE FRACTURES : BONEVIEW®

Notre étude et les interprétations radiologiques qui en découlent, se basent sur l'utilisation d'un logiciel d'IA d'aide à la détection des fractures mis à notre disposition par l'entreprise française Gleamer, nommé Boneview®.
L'ensemble de données d'entraînement comprenait 15% de patients âgés de moins de 18 ans.

Le logiciel Boneview®, d'aide à la détection de fracture, a été développé à partir d'un ensemble de données provenant de 312 602 radiographies de patients souffrants de traumatismes. Ces patients ont été hospitalisés dans plus de 60 services de radiologie privés français, de janvier 2011 à mai 2021. Sur la base de cet échantillon de données, divisé de manière aléatoire en 70% d'ensemble d'entraînement, 10% de validation et 20% de tests internes, un réseau neuronal de *deep-learning,* basé sur le framework "Detectron 2" (7), a été conçu, entraîné, optimisé et validé pour détecter et localiser les fractures sur radiographies

[1] Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016, relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données, et abrogeant la directive 95/46/CE (règlement général sur la protection des données). Également nommé Règlement Général sur la Protection des Données (RGPD).

numériques à résolution native. L'ensemble de données d'entraînement comprenait 15% de patients âgés de moins de 18 ans.

Ce système d'IA met en évidence chaque région d'intérêt par une case et fournit un score de confiance concernant l'existence d'une fracture dans la région d'intérêt.

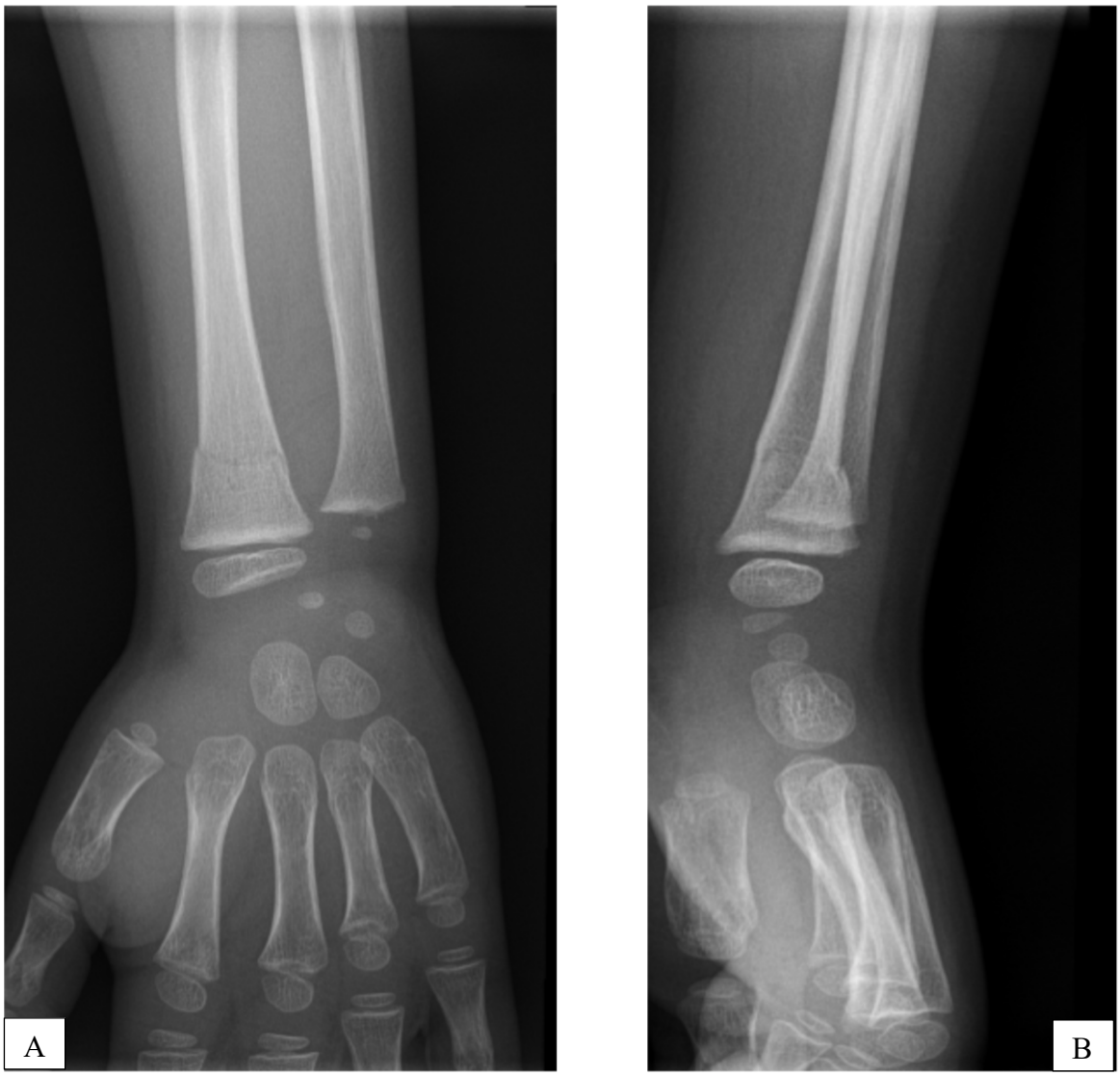

Figure 1 : Exemple de radiographies initiales du poignet face (A) et profil (B)

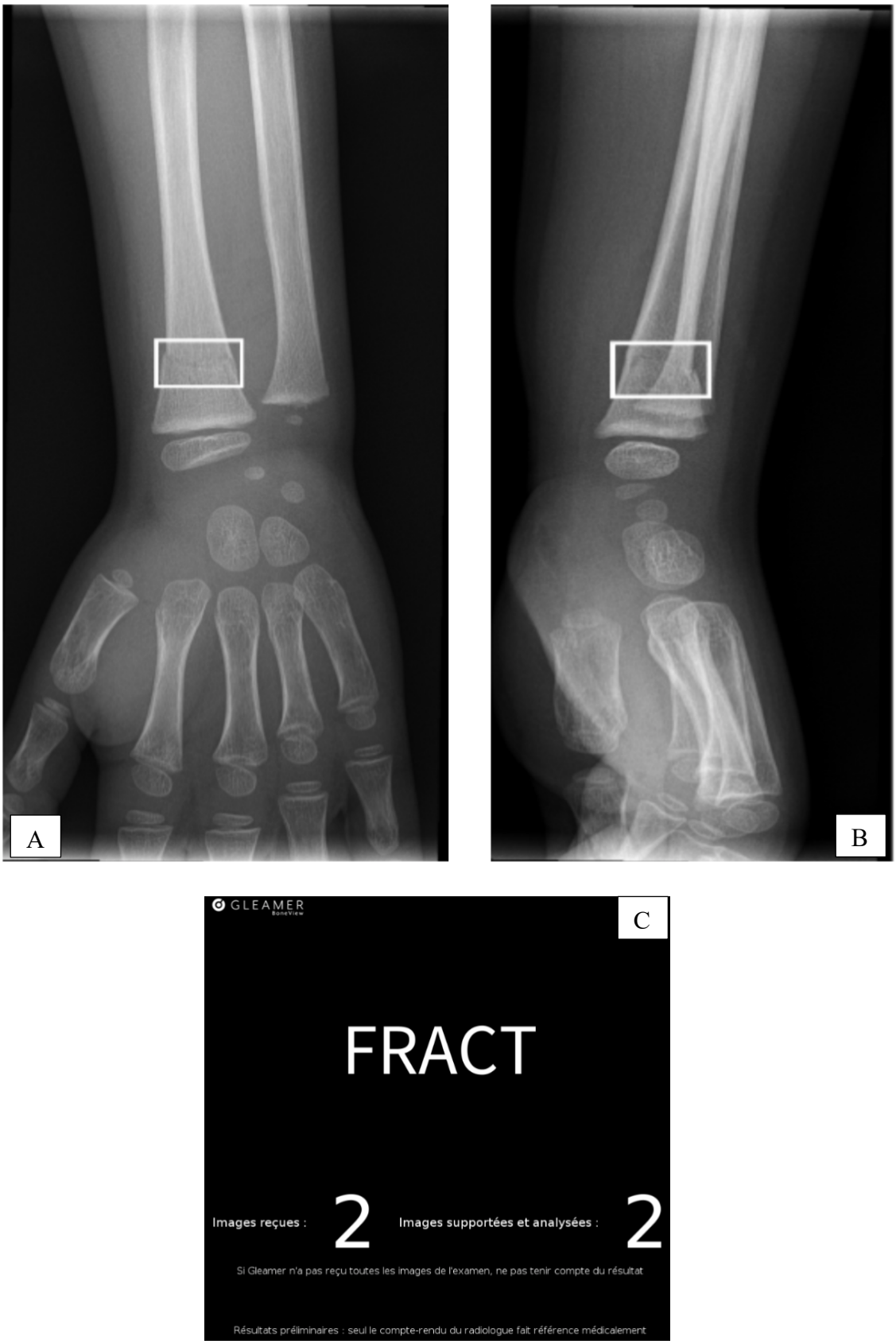

Figure 2 : Exemple de radiographies du poignet face (A) et profil (B) après identification par le logiciel Boneview® des régions d'intérêt et capture secondaire envoyée sur le PACS (C).

Ladite intelligence artificielle affiche l'information « fracture » dès lors que la certitude diagnostique avoisine les 90 % ou « doute fracture » lorsque la certitude diagnostique est comprise entre 70 et 90 %.

MATERIEL UTILISE POUR LES RADIOGRAPHIES

Les tables de radiographie utilisées pour réaliser les examens des patients venant consulter au service d'accueil des urgences pédiatriques de l'Hôpital Nord de Marseille, sont des tables PHILIPS® capteur plan. Les incidences en cas de traumatisme sont acquises en face et profil. Aucune incidence oblique n'a été réalisée chez les patients de notre étude et aucun cliché comparatif n'a été réalisé.

POPULATION ETUDIEE ET CRITERES D'INCLUSION/EXCLUSION DES PATIENTS DANS L'ETUDE

L'échantillon de 400 examens radiographiques, correspondant à 804 radiographies, utilisé au cours de l'étude, a été construit rétrospectivement à partir du 31 janvier 2020. Une recherche ante-chronologique, dans la base de données d'examens radiologiques, a permis de réunir 100 patients consécutifs, ayant fait l'objet d'une exploration radiographique lors d'une consultation aux urgences pédiatriques à l'hôpital Nord de Marseille, pour l'examen des articulations du coude, du poignet, du genou et de la cheville.
L'ensemble des radiographies sélectionnées a été pris à un moment (t) unique pour chaque patient, moment correspondant à la consultation initiale aux urgences.

Les patients sélectionnés devaient avoir moins de 16 ans, avoir été admis au service d'accueil des urgences pédiatriques de l'hôpital Nord de Marseille pour traumatisme et avoir bénéficié d'un examen radiographique d'une des quatre articulations précitées.

Les patients présentant une incertitude sur leur âge ou un historique clinique flou, avec une notion traumatique incertaine, n'ont pas été retenus. Le suivi du patient après le traumatisme a été vérifié grâce au dossier patient informatisé (Axigate ®) afin d'exclure les dossiers dont l'évolution clinique n'était pas en faveur du diagnostic initial.

Les radiographies de contrôle « post-réduction d'une luxation » ou « réfection de plâtre » ont également été exclues de l'analyse de la présente étude.

ANONYMISATION DES DONNEES A CARACTERE PERSONNELLE

Dans le but de garantir une anonymisation des données relatives aux patients, deux étapes de codification ont été nécessaires. La première étape a associé le nom du patient à un numéro d'hospitalisation. Ledit numéro d'hospitalisation a, par la suite, été remplacé par un numéro, créer aléatoirement, non rattachable au nom du patient. Cette codification aléatoire a permis une interprétation anonymisée de l'ensemble des examens inclus dans l'étude tel que le prévoit la loi sur la protection des données personnelles.

LECTURE ET INTERPRETATION DES IMAGERIES MEDICALES

L'ensemble des 400 examens radiographiques, recueilli dans le cadre de l'étude, a été interprété par un radiologue senior spécialisé en radiopédiatrie ayant plus de 20 ans d'expérience (Pr. K.Chaumoître et Pr. M.Panuel). Ces imageries ont également fait l'objet d'une interprétation par un Interne de radiologie en fin de cursus (M. A. Planche). Les lecteurs étaient aveugles aux comptes rendus des autres radiologues impliqués ainsi qu'au diagnostic formulé par l'IA.
L'ensemble des interprétations a été réalisé sur des consoles de radiologie dédiées, sans limite de temps et avec l'accès au « bon d'examen », comportant les données cliniques du patient, rédigées par le médecin urgentiste.

Il doit être notifié, qu'afin d'être en adéquation avec les capacités de détection du logiciel Boneview®, que la recherche d'épanchement sur les imageries recueillies a été secondairement exclue. Un tel épanchement ne pouvant être détecté par l'IA que dans la région du coude, sa prise en compte aurait entravé l'homogénéité des résultats.

Par la suite, les imageries médicales ont été extraites du serveur de l'Assistance Public Hôpitaux de Marseille sous format DICOM[2], anonymisées et envoyées, de manière sécurisée, à la société Gleamer. Après traitement de ces examens, par le logiciel d'aide à la détection de fracture Boneview®, un envoi sécurisé des résultats obtenus à destination de l'Interne de radiologie a été effectué.

[2] Le format *Digital Imaging and Communication in Medicine* désigne la norme relative au format des fichiers numériques créés lors d'examens d'imagerie médicale.

Ultérieurement, l'Interne a réalisé une nouvelle interprétation, en aveugle de sa précédente analyse, de l'ensemble des examens d'imagerie recueillis avec l'aide du logiciel d'intelligence artificielle Boneview®.

Après cette seconde interprétation, l'Interne n'a modifié son compte-rendu dans le seul cas où, le logiciel Boneview® avait décelé une fracture non visualisée lors de sa première interprétation. A contrario, les comptes rendus n'ont fait l'objet d'aucune normalisation dès lors que l'IA signifiait une absence de fracture. Un défaut de confiance dans les performances de l'intelligence artificielle, notamment dans sa spécificité, a conduit l'Interne à ne prendre en considération que les diagnostics dits « en excès ».

Les comptes rendus résultants desdites interprétations, rédigés par l'Interne et le senior, ont été classés de manière binaire en deux catégories : « présence d'une fracture ou doute sur l'existence d'une fracture » versus « pas de fracture visible sur l'imagerie ». La même dichotomie catégorielle a servi de classification à l'interprétation des imageries faites par le logiciel Boneview®.

ANALYSE STATISTIQUE DES DONNEES RECUEILLIES

En considérant l'interprétation du Professeur universitaire comme référence « senior », les données recueillies lors de l'étude ont fait l'objet d'une analyse statistique.

Les sensibilités, les spécificités, les valeurs prédictives positives et négatives des interprétations par articulation, effectuées par l'Interne avec ou sans l'appui de l'intelligence artificielle, ont été évaluées.

La sensibilité d'un test mesure sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée. Elle s'oppose à la spécificité, qui mesure la capacité d'un test à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée.
La validité prédictive est quant à elle explorée par la valeur prédictive positive qui correspond à la probabilité que la maladie soit présente lorsque le test est positif. La valeur prédictive négative est, à contrario, la probabilité que la maladie ne soit pas présente lorsque le test est négatif.

Afin d'être comparé, l'ensemble de ces valeurs ont fait, par la suite, l'objet d'un test de Student (t) apparié avec un seuil de significativité à 5%.

En outre, les données d'interprétation, ont été analysées par un test de concordance-diagnostique dit test non paramétrique Kappa de Cohen (K). Ce test permet de chiffrer l'accord entre deux ou plusieurs observateurs ou techniques lorsque les jugements sont qualitatifs.
L'estimation de l'accord entre les jugements catégoriels appliqués aux mêmes objets, fournis par deux ou plusieurs observateurs ou techniques, prend en compte la concordance aléatoire et permet ainsi de s'en affranchir.

La « concordance » signifie la proportion de sujets pour lesquels il y a accord entre les observateurs (22). On évalue ladite concordance grâce au calcul d'un coefficient nommé *kappa*. Ce coefficient varie de 0 à 1 en fonction du niveau de concordance observé.

Le tableau, ci-après, expose l'interprétation généralement admise du coefficient *kappa*. Cette interprétation, basée sur une étude unique et subjective (23), reste cependant discutable.

Tableau 1 : Interprétation du coefficient Kappa selon Landis et al. (23)

Coefficient Kappa	**Estimation du degré de concordance**
0	Mauvais
0 à 0.2	Négligeable
0.2 à 0.4	Faible
0.4 à 0.6	Moyen
0.6 à 0.8	Bon
0.8 à 1	Excellent

RESULTATS

CARACTERISTIQUES DE LA POPULATION ETUDIEE

Au cours de l'interprétation des quatre séries de 100 examens radiographiques pédiatriques traumatiques des principales articulations, par le *gold standard*, représenté par le senior, 242 examens ne présentaient ni fracture ni épanchement, 124 présentaient une fracture et 34 uniquement un épanchement. Afin d'être en adéquation avec les capacités de détection du logiciel Boneview®, les données relatives aux épanchements ont été exclues de la suite de l'analyse.

Ces résultats, présentés dans le Tableau 2, mettent en évidence, en fonction de l'articulation explorée, une forte disparité des taux d'examens présentant une anomalie dite traumatique.

Seuls 9 % des examens intéressants un genou traumatique sont pathologiques contre 25 % concernant les chevilles traumatiques, 36 % concernant les coudes traumatiques et 54 % concernant les poignets traumatiques.

Tableau 2 : Résumé des interprétations des imageries en fonction du site étudié

Localisation	Pas de fracture / épanchement	Épanchement sans fracture	Fracture	Nombre d'examens
Coude	54	10	36	100
Poignet	46	0	54	100
Genou	75	16	9	100
Cheville	67	8	25	100
Total	242	34	124	400
% Total	60,5%	8,5 %	31%	100%

L'âge moyen des 400 patients inclus était de 9,44 ans avec une fourchette d'âge entre 11 mois et 16 ans et une médiane d'âge à 10 ans.

Concernant le sex-ratio, 237 des 400 patients (59,25%) étaient des garçons et seulement 163 patients étaient des filles (40,75%). Les garçons étaient en moyenne plus âgés que les filles avec un âge moyen de 9,96 ans contre 8,74 ans.

Tableau 3 : Résumé des caractéristiques démographiques de la population étudiée

Localisation	**Moyenne d'âge**	**Moyenne d'âge ♀**	**Moyenne d'âge ♂**	**Médiane d'âge**	**Fourchette d'âge (an)**	**Nombre de ♀**	**Nombre de ♂**
Coude	7.1 ans	6,1 ans	8 ans	6 ans	[1-16]	43	57
Poignet	9.3 ans	8,2 ans	9,9 ans	10 ans	[1-15]	36	74
Genou	11.3 ans	11,4 ans	11,2 ans	12 ans	[2-16]	33	77
Cheville	10 ans	9,4 ans	10,64 ans	11 ans	[0,92-16]	51	49
Total	9.4 ans	8,7 ans	9,9 ans	10 ans	[0,92-16]	163	237

INTERPRETATIONS RADIOLOGIQUES, ANALYSE PAR ARTICULATION

- ***Coude***

Avec 7 fractures non visualisées et 2 fractures décelées à tort (**Tableau *4***), l'interprétation radiologique effectuée par l'Interne, sans le concours de l'IA, affiche, en référence au *gold standard*, une sensibilité de 0,81 associée à une spécifié évaluée à 0,97 (**Tableau *5***).

D'autre part, l'interprétation des examens d'imagerie par le logiciel d'IA Boneview® affiche une discordance en défaut moins élevée, avec seulement 5 examens radiographiques faussement négatifs, et un taux de faux positifs plus élevé avec 8 examens discordants par excès (**Tableau *4***).

En outre, l'association avec l'IA a permis à l'Interne de radiologie d'améliorer la valeur prédictive négative et la sensibilité de ses interprétations (**Tableau *5***). La sensibilité atteint dès lors 0,92 et 4 examens faussement négatifs, en première lecture, ont pu être redressés. Toutefois, avec 7 examens supplémentaires positivés à tort, la valeur prédictive positive et la spécificité des interprétations décroit passant de 0,97 à 0,86 (**Tableau *5***).

La concordance κ a quant à elle été améliorée, avec l'aide de l'IA, passant de 0,80 à 0,93 (**Tableau *4***).

Tableau 4 : Comparatif des interprétations radiologiques du coude entre le senior, l'interne et l'IA

		Senior	**Interne**	**IA**	**Interne + IA**
Nombre d'examens radiographiques	Fracture	36	31	42	37
	Pas de fracture	64	69	58	63
	Discordance en excès		2	8	9
	Discordance en défaut		7	5	3
Concordance diagnostique (K)	Senior		0,80	0,71	0,93
	Interne			0,66	

IA : logiciel Boneview®.

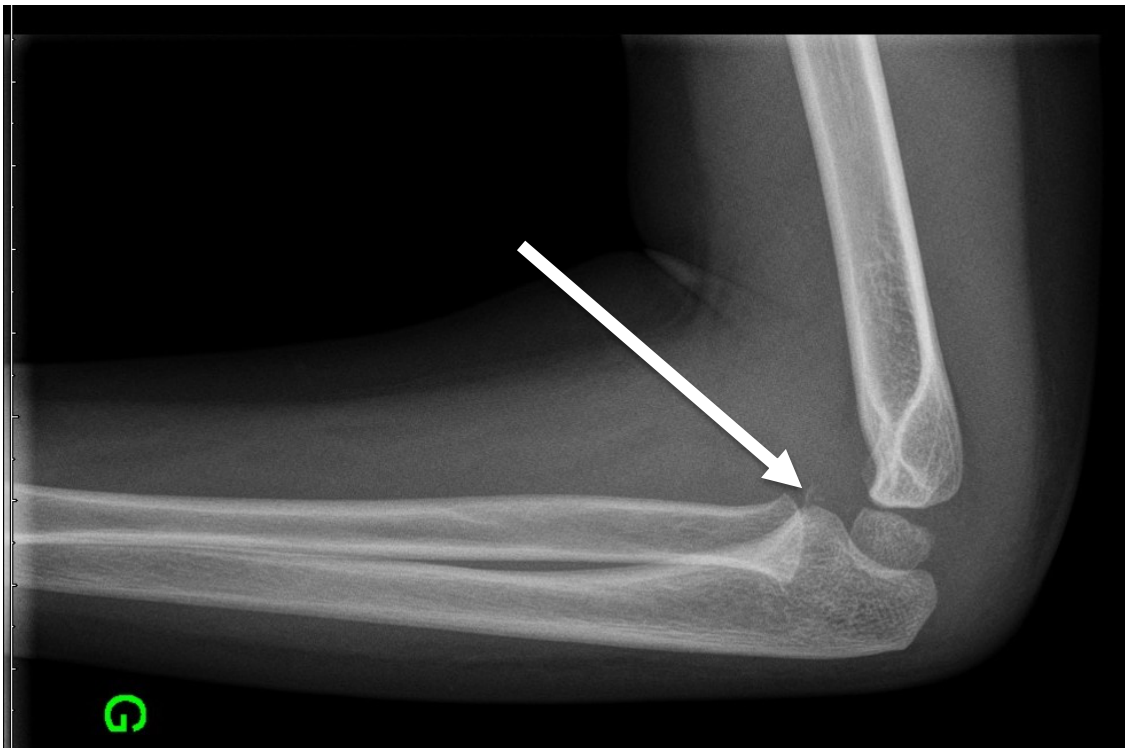

Figure 3 : Exemple d'une radiographie du coude gauche, de profil, avec une fracture du processus coronoïde (flèche) visualisée par l'Interne et non détectée par le logiciel d'IA.

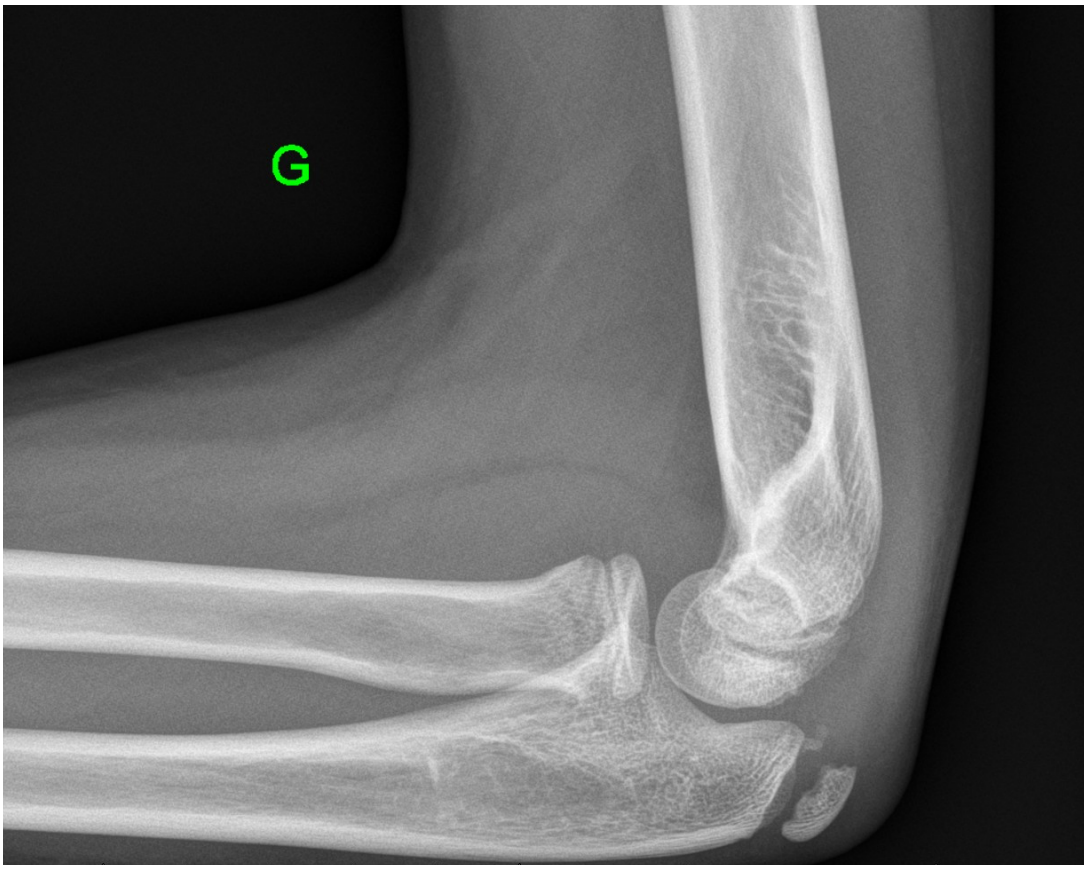

Figure 4 : Exemple d'une radiographie du coude gauche, de profil, avec une fracture de la métaphyse radiale non visualisée par l'Interne et détectée par le logiciel d'IA

Tableau 5 : Tableau de contingence et données de performances diagnostiques relative à l'articulation du coude

Conclusions d'examens			**Sensibilité**	**Spécificité**	**VPP**	**VPN**
	Senior +	**Senior -**				
Interne +	29	2	0,80	0,97	0,93	0,90
Interne -	7	67				
Interne/IA +	34	9	0,92	0,86	0,79	0,95
Interne/IA -	3	54				
IA +	34	8	0,87	0,87	0,81	0,91
IA -	5	53				

- ***Poignet***

Avec 7 fractures non visualisées et seulement 1 fracture décelée à tort (**Tableau *6***), l'interprétation radiologique effectuée par l'Interne, sans le concours de l'IA, affiche, en référence au *gold standard*, une sensibilité de 0,86 associée à une spécifié évaluée à 0,98 (**Tableau *7***).

D'autre part, l'interprétation des examens d'imagerie par le logiciel d'IA Boneview® affiche une discordance en défaut moins élevée, avec seulement 1 examen radiographique faussement négatif, et un taux de faux positif plus élevé avec 2 examens discordants par excès (**Tableau *6***).

En outre, l'association avec l'IA a permis à l'Interne de radiologie d'améliorer la valeur prédictive négative et la sensibilité de ses interprétations (**Tableau *7***). La sensibilité atteint dès lors 0,98 et 6 examens faussement négatifs, en première lecture, ont pu être redressés. Toutefois, avec 7 examens supplémentaires positivés à tort, la valeur prédictive positive et la spécificité des interprétations décroit passant de 0,98 à 0,94 (**Tableau *7***).

La concordance κ a quant à elle été améliorée, avec l'aide de l'IA, passant de 0,86 à 0,94 (**Tableau *6***).

Tableau 6 : Comparatif des interprétations radiologiques du poignet entre le senior, l'Interne et l'IA

		Senior	**Interne**	**IA**	**Interne + IA**
Nombre d'examens radiographiques	Fracture	54	49	49	53
	Pas de fracture	46	51	51	47
	Discordance en excès		1	2	3
	Discordance en défaut		7	1	1
Concordance diagnostique (K)	Senior		0,86	0,94	0,94
	Interne			0,84	

IA : logiciel Boneview®.

Figure 5 : Exemple d'une radiographie du poignet droit, de profil, avec une fracture en motte de beurre de la métaphyse radiale non visualisée par l'Interne et détectée par l'IA.

Tableau 7 : Tableau de contingence et données de performances diagnostiques relatives à l'articulation du poignet

Conclusions d'examens			**Sensibilité**	**Spécificité**	**VPP**	**VPN**
	Senior +	**Senior -**				
Interne +	42	1	0,86	0,98	0,98	0,88
Interne -	7	50				
Interne/IA +	52	3	0,98	0,94	0,94	0,98
Interne/IA -	1	44				
IA +	47	2	0,98	0,96	0,96	0,98
IA -	1	50				

- ***Genou***

Avec seulement une fracture non visualisée et 4 fractures décelées à tort (**Tableau *8***), l'interprétation radiologique effectuée par l'Interne, sans le concours de l'IA, affiche, en référence au *gold standard*, une sensibilité de 0,88 associée à une spécifié évaluée à 0,96 (**Tableau *9***).

D'autre part, l'interprétation des examens d'imagerie par le logiciel d'IA Boneview® affiche une discordance en défaut plus élevée, avec 4 examens radiographiques faussement négatifs, et un faible taux de faux positif avec seulement 1 examen discordant par excès (**Tableau *8***).

En outre, l'association avec l'IA a permis à l'Interne d'améliorer la valeur prédictive négative et la sensibilité de ses interprétations (**Tableau *9***). La sensibilité atteint dès lors la valeur idéale d'1, le seul examen faussement négatif, en première lecture, ayant pu être décelé. Toutefois, avec 1 examen supplémentaire positivé à tort, la valeur prédictive positive et la spécificité des interprétations décroit passant de 0,96 à 0,94 (**Tableau *9***).

La concordance κ a quant à elle été améliorée, avec l'aide de l'IA, passant de 0,73 à 0,76 (**Tableau *8***).

Tableau 8 : Comparatif des interprétations radiologiques du genou entre le senior, l'Interne et l'IA

		Senior	**Interne**	**IA**	**Interne + IA**
Nombre d'examens radiographiques	Fracture	9	12	4	14
	Pas de fracture	91	88	96	86
	Discordance en excès		4	1	5
	Discordance en défaut		1	6	0
Concordance diagnostique (K)	Senior		0,73	0,43	0,75
	Interne			0,20	

IA : logiciel Boneview®.

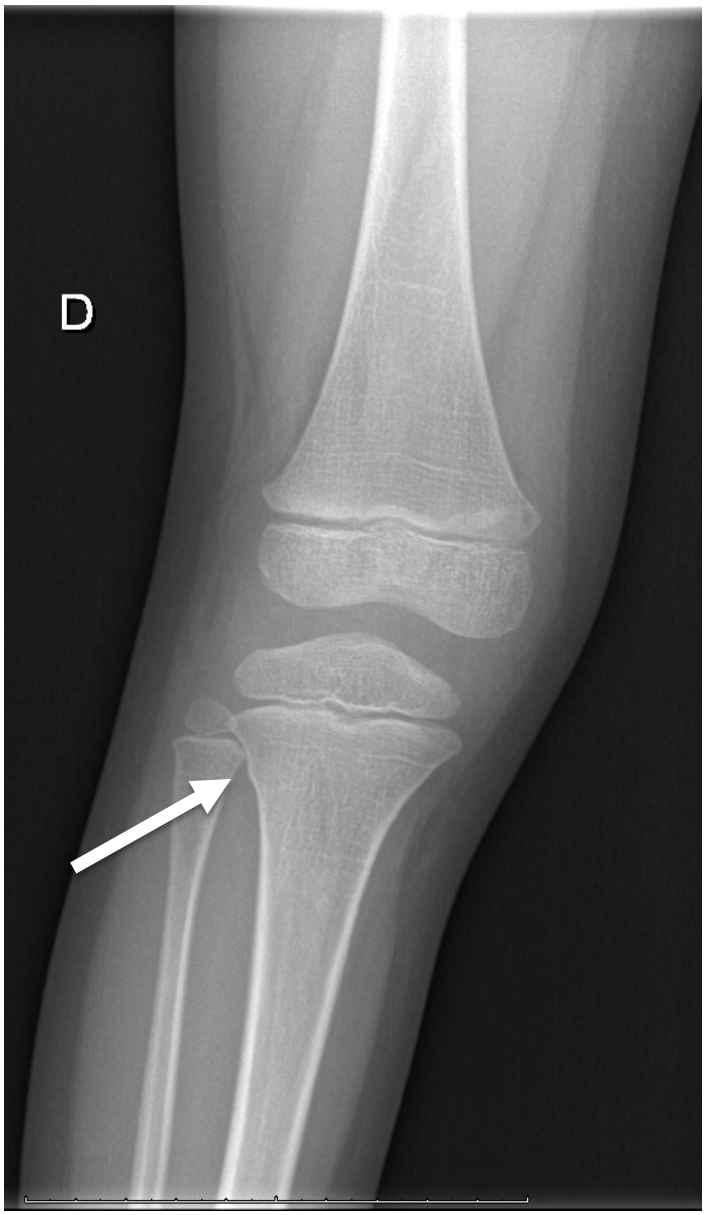

Figure 6 : Exemple d'une radiographie du genou droit, de face, avec une fracture en motte de beurre de la métaphyse tibiale (flèche) non visualisée par l'Interne et détectée par la suite par le logiciel d'IA

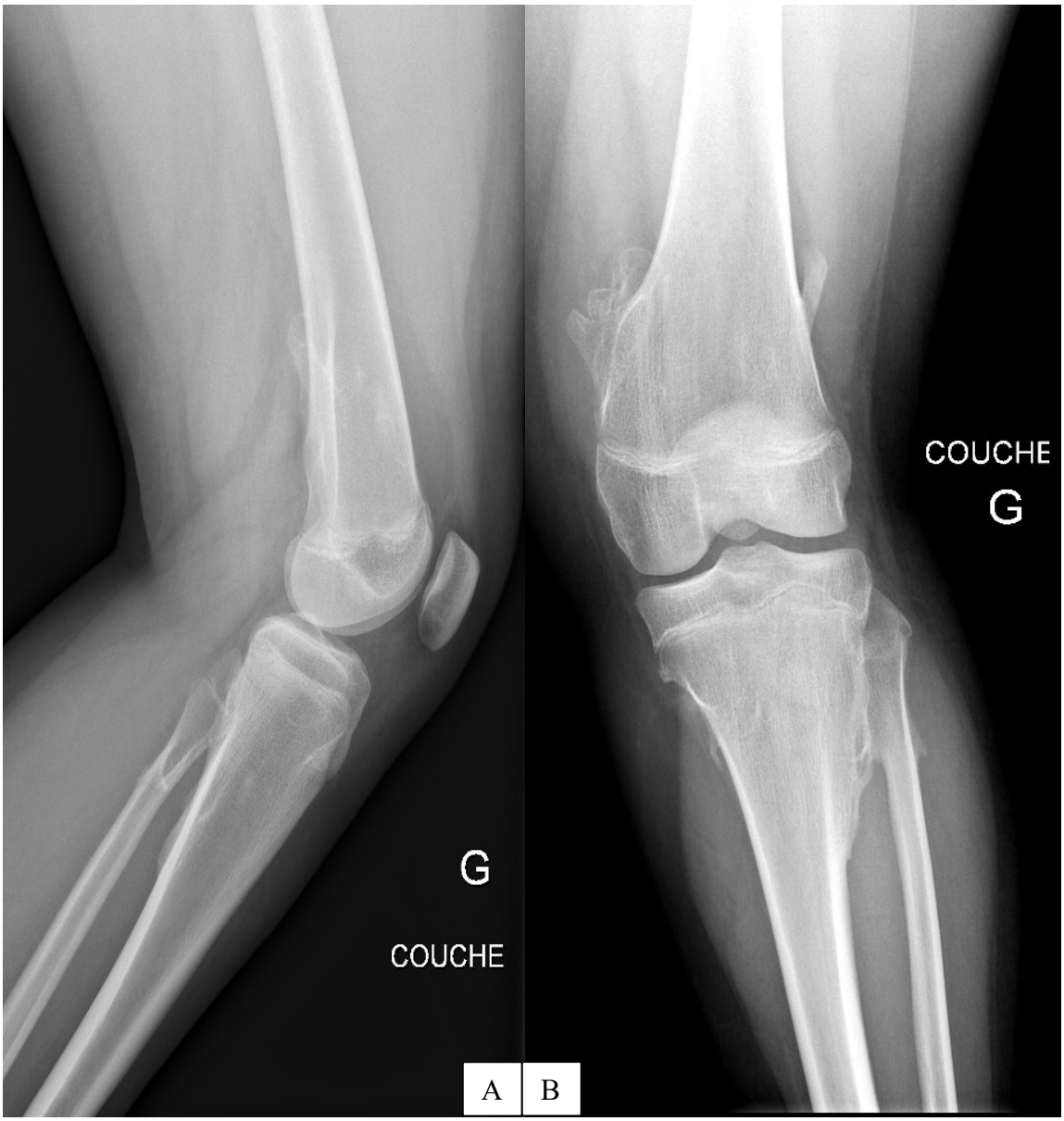

Figure 7 : Exemple d'une radiographie d'un genou gauche de profil (A) et de face (B) avec une maladie des exostoses multiples, détectée à tort par l'IA comme des fractures métaphysaires

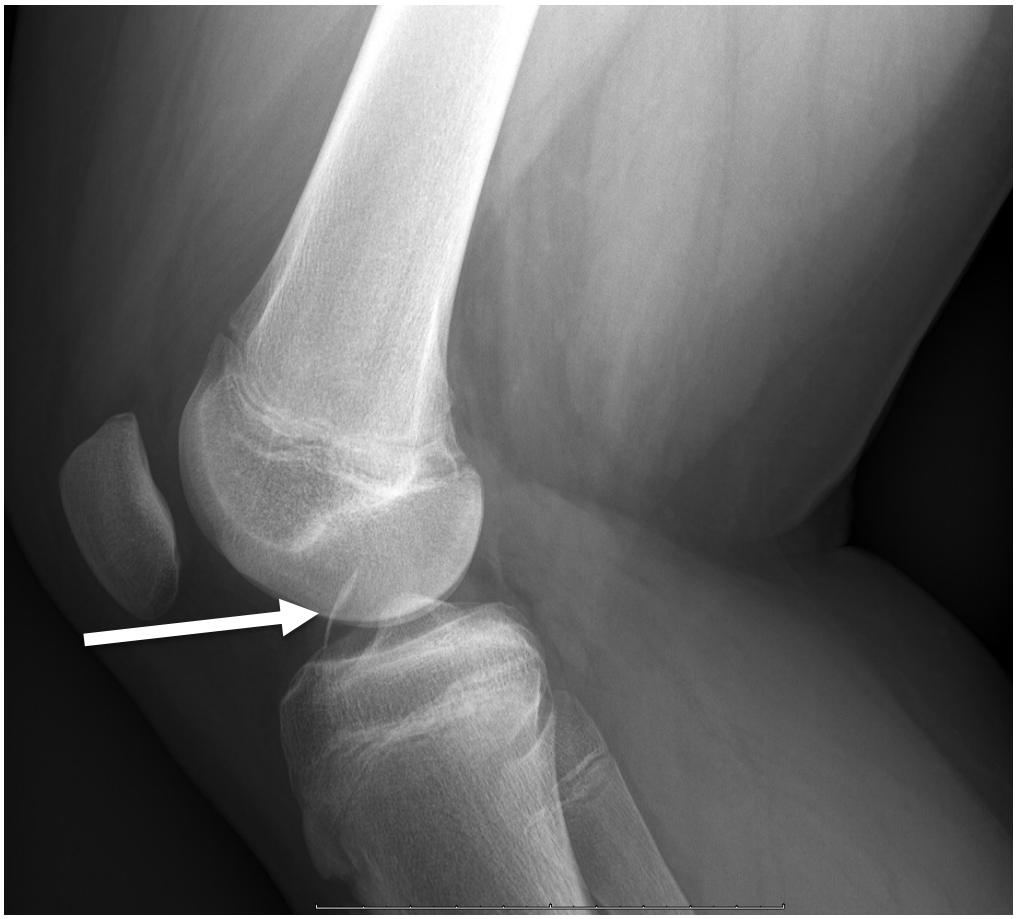

Figure 8 : Exemple d'une radiographie d'un genou gauche, de face, avec un fragment ostéo-chondrale au sein de l'échancrure inter-condylienne (flèche) visualisée par l'Interne et non décelée par le logiciel d'IA

Tableau 9 : Tableau de contingence et données de performances diagnostiques relatives à l'articulation du genou

Conclusions d'examens			**Sensibilité**	**Spécificité**	**VPP**	**VPN**
	Senior +	**Senior -**				
Interne +	8	4	0,89	0,96	0,67	0,99
Interne -	1	87				
Interne/IA +	9	5	1	0,94	0,64	1
Interne/IA -	0	86				
IA +	3	1	0,33	0,99	0,75	0,94
IA -	6	90				

- ***Cheville***

Avec 7 fractures non visualisées et 7 fractures décelées à tort (**Tableau *10***) l'interprétation radiologique effectuée par l'Interne, sans le concours de l'IA, affiche, en référence au *gold standard*, une sensibilité de 0,73 associée à une spécifié évaluée à 0,90 (**Tableau *11***).

D'autre part, l'interprétation des examens d'imagerie par le logiciel d'IA Boneview® affiche une discordance en défaut plus élevée, avec 10 examens radiographiques faussement négatifs, et un taux de faux positif équivalent avec 7 examens discordants par excès (**Tableau *10***).

En outre, l'association avec l'IA a permis à l'Interne d'améliorer la valeur prédictive négative et la sensibilité de ses interprétations (**Tableau *11***). La sensibilité atteint dès lors 0,85 et 3 examens faussement négatifs, en première lecture, ont pu être redressés. Toutefois, avec 5 examens supplémentaires positivés à tort, la valeur prédictive positive et la spécificité des interprétations décroit passant de 0,90 à 0,83 (**Tableau *11***).

La concordance κ a quant à elle été améliorée, avec l'aide de l'IA, passant de 0,66 à 0,69 (**Tableau *10***).

Tableau 10 : Comparatif des interprétations radiologiques de la cheville entre le senior, l'Interne et l'IA

		Senior	**Interne**	**IA**	**Interne + IA**
Nombre d'examens radiographiques	Fracture	25	26	26	27
	Pas de fracture	75	74	74	73
	Discordance en excès		7	7	12
	Discordance en défaut		7	10	4
Concordance diagnostique (K)	Senior		0,66	0,47	0,69
	Interne			0,66	

IA : logiciel Boneview®.

Figure 9 : Exemple d'une radiographie de la cheville droite, de face, avec fracture ostéo-chondrale de la pointe de la fibula (flèche) non visualisée par l'Interne et détectée par le logiciel d'IA

Tableau 11 : Tableau de contingence et données de performances diagnostiques relatives à l'articulation de la cheville

Conclusions d'examens			Sensibilité	Spécificité	VPP	VPN
	Senior +	Senior -				
Interne +	19	7	0,73	0,90	0,73	0,90
Interne -	7	67				
Interne/IA +	23	12	0,85	0,83	0,66	0,94
Interne/IA -	4	61				
IA +	19	7	0,65	0,90	0,73	0,86
IA -	10	64				

INTERPRETATIONS RADIOLOGIQUES, ANALYSE GENERALE DES CONCORDANCES DIAGNOSTIQUES

- ***Senior et Interne, sans le concours de l'intelligence artificielle, analyse des concordances diagnostiques***

Les résultats de la présente étude montrent une concordance diagnostique moyenne, entre l'Interne et le senior, de 0,76 avec une faible disparité en fonction des différentes articulations étudiées.

Bien que la disparité des résultats soit faible, la concordance diagnostique présente un meilleur score lorsqu'il s'agit de l'articulation du poignet avec un résultat de 0,86. Comme le montre le tableau de résultats ci-dessous, ce score de concordance, est diminué dès lors qu'il concerne les articulations du coude, du genou et de la cheville pour être respectivement de 0,80, 0.73 et 0,66.

Tableau 12 : Concordances diagnostiques entre l'Interne et le Senior

	Coude	Poignet	Genou	Cheville	Moyenne
Concordance diagnostique (kappa)	0.80	0.86	0.73	0.66	**0,76**

- ***Interne et IA, analyse des concordances diagnostiques***

À contrario des résultats précédemment exposés, dès lors qu'il s'agit de mettre en parallèle l'interprétation radiologique faite par l'Interne et l'interprétation proposée par le logiciel Boneview® d'IA d'aide à la détection des fractures, le score de concordance diagnostique moyen chute.

Le Tableau 13, ci-dessous, retranscrit un score de concordance diagnostique moyen, sur l'ensemble des 400 examens interprétés par l'Interne de radiologie et le logiciel Boneview®, de 0.59 avec un score maximal de 0.84, pour l'articulation du poignet, et un score minimal de 0.20 pour l'articulation du genou. Dans cet intervalle, les interprétations relatives aux articulations de la cheville et du coude donnent des scores de concordance-diagnostique mesurés de 0.66.

Tableau 13 : Concordances diagnostiques entre l'Interne et l'IA

	Coude	Poignet	Genou	Cheville	Moyenne
Concordance diagnostique (kappa)	0.66	0.84	0.20	0.66	**0,59**

- ***Senior et IA, analyse des concordances diagnostiques***

Parallèlement, les résultats de la présente étude montrent une concordance diagnostique moyenne, entre le senior et le logiciel Boneview®, légèrement supérieure à celle précédent citée, concernant l'Interne et ledit logiciel, pour atteindre un score de 0.64.

Dans les résultats décrits dans le Tableau 14, il est à souligner une forte disparité en fonction des différentes articulations étudiées. Les interprétations des imageries relatives à l'articulation du poignet présentent une concordance diagnostique quasi parfaite avec un score atteignant 0.94 tandis que le score relatif à l'articulation du genou atteint péniblement la valeur de 0.43. Ce score de concordance est modifié dès lors qu'il concerne les articulations de la cheville et du coude, pour être respectivement de 0.47 et 0.71.

Tableau 14 : Concordances diagnostiques entre le senior et l'IA

	Coude	**Poignet**	**Genou**	**Cheville**	**Moyenne**
Concordance diagnostique (kappa)	0.71	0.94	0.43	0.47	**0,64**

Au vu des résultats précédemment énoncés, la concordance diagnostique moyenne du logiciel d'intelligence artificielle avec le senior reste inférieure à la concordance diagnostique entre l'Interne et le senior. L'articulation du poignet fait cependant exception. Le logiciel Boneview® présente une concordance de 0.94 avec le Professeur contre 0.86 seulement entre l'Interne et le Professeur.

- ***Senior et Interne, avec le concours de l'intelligence artificielle, analyse des concordances diagnostiques***

Après une seconde interprétation par l'Interne, aidé du logiciel Boneview®, des 400 examens radiographiques recueillis, les résultats de la présente étude montrent une concordance diagnostique moyenne, entre l'Interne et le senior, de 0.83. Ce score met en évidence une amélioration de la concordance des diagnostics posés par l'Interne et le senior, comparativement à une interprétation des imageries sans le concours du logiciel d'IA.

L'amélioration la plus franche intéresse l'articulation du coude avec une augmentation du coefficient de concordance passant de 0.80 à 0.93. Le coefficient de concordance relatif à l'articulation du poignet présente également un accroissement passant de 0.86 à 0.94. En revanche, les coefficients de concordance diagnostique des articulations de la cheville et du genou montrent une progression plus modeste évoluant respectivement de 0.66 à 0.69 et de 0.73 à 0.76.

Tableau 15 : Concordances diagnostiques entre l'Interne avec l'IA et le senior

	Coude	Poignet	Genou	Cheville	Moyenne
Concordance diagnostique (kappa)	0.93	0.94	0.76	0.69	**0,83**

Il est à noter que les modifications d'interprétation, réalisées par l'Interne lors de la seconde visualisation des imageries avec le concours du logiciel Boneview®, ont uniquement pris en considération l'adjonction, dans les conclusions diagnostiques, de lésions traumatiques non détectées lors de la première interprétation. Aucune correction du compte-rendu n'a été effectuée, dès lors que le logiciel d'intelligence artificielle ne détectait pas, à contrario de l'Interne, une lésion traumatique.

SIGNIFICATIVITE STATISTIQUE DE LA MODIFICATION DES PERFORMANCES ET DE LA CONCORDANCE DIAGNOSTIQUE

Les résultats de l'étude mettent en évidence une amélioration significative de la sensibilité des interprétations effectuées par l'Interne avec le concours du logiciel d'IA Boneview®. La sensibilité moyenne passant de 0.82 à 0.94 avec une différence absolue de 0.12 (**$p < 0.05$** ; [+0.1271 ; +0.1076]). Sans significativité statistique, la valeur prédictive négative s'en trouve elle aussi améliorée. A contrario, lesdites interprétations accusent une baisse, non significative, de leur spécificité et de leur valeur prédictive positive.

Concernant la concordance diagnostique moyenne, une amélioration non statistiquement significative est également observée.

Ainsi, le logiciel d'IA permet à l'Interne d'améliorer ses performances diagnostiques avec une meilleure détection des fractures appendiculaires de l'enfant sans impacter de manière significative sa spécificité, sa VPP ou sa VPN.

Tableau 16 : Concordance et performance diagnostiques de l'Interne avec et sans le concours de l'IA

	Interne sans IA	**Interne et IA**	**Différence absolue (p-value)**	**Intervalle de confiance à 95%**
Sensibilité moyenne	0,82	0,94	0.1174 **(p < 0.05)**	[+0.1271 ; +0.1076]
Spécificité moyenne	0,95	0,89	0.0598 (p = 0.07)	[-0.0093 ; +0.1288]
Valeur prédictive positive moyenne	0.83	0.76	0.0684 (p = 0.09)	[-0.0198 ; +0.1565]
Valeur prédictive négative moyenne	0.92	0.97	0.0468 (p = 0.09)	[-0.0198 ; +0.1565]
Concordance diagnostique moyenne	0.76	0.83	0.062 (p = 0.12)	[-0.1548 ; 0.031]

En outre, la comparaison entre les performances diagnostiques de l'Interne et de l'IA montre, pour l'ensemble des indicateurs étudiés, une supériorité non statistiquement significative de l'Interne seul par rapport au logiciel d'IA. Il en est de même pour la concordance avec le senior qui est de 0.76 pour l'Interne, contre seulement 0.64 pour l'IA.

Tableau 17 : Concordances et performances diagnostiques comparées entre l'Interne et l'IA

	Interne sans IA	**IA**	**Différence absolue (p-value)**	**Intervalle de confiance à 95%**
Sensibilité moyenne	0,82	0,71	0.1107(p = 0.52)	[-0.3794 ; 0.6008]
Spécificité moyenne	0,95	0,93	0.02 (p = 0.48)	[-0.0678 ; 0.1139]
Valeur prédictive positive moyenne	0.83	0.81	0.015 (p = 0.75)	[-0.1219 ; 0.152]
Valeur prédictive négative moyenne	0.92	0.92	0.005(p = 0.90)	[-0.117 ; 0.1071]
Concordance diagnostique moyenne	0.76	0.64	0.21(p = 0.22)	[-0.1343 ; 0.3828]

DISCUSSION

La multitude de consultations pour suspicion de fracture, au sein des services des urgences pédiatriques, et les nombreuses erreurs de prise en charge qui en découlent font du diagnostic des fractures pédiatriques des membres un enjeu d'importance. Or, la carence que connaît le territoire français en radiologues spécialisés en pédiatrie, associée à la complexité diagnostique des fractures de l'enfant, font de cet enjeu de santé publique une problématique difficilement solvable.

L'aide d'un logiciel d'intelligence artificielle est une piste d'amélioration des performances diagnostiques des radiologues généralistes, amélioration des performances qui pourrait partiellement pallier à la faible démographie en radiologues spécialisés en pédiatrique et, ainsi, parfaire la prise en charge des populations juvéniles sur le territoire national.

Cette étude, démontre que l'utilisation d'un logiciel d'intelligence artificielle d'aide au diagnostic, par un Interne de radiologie en fin de cursus, assimilable à un radiologue généraliste, lui permet d'améliorer significativement sa performance diagnostique, avec comme *gold standard* un Professeur de radiopédiatrie expérimenté, lorsqu'il s'agit de la détection de fractures sur des radiographies traumatiques pédiatriques. Sa sensibilité de détection passe de 0.82 à 0,94 (intervalle de confiance à 95% [+0.1271 ; +0.1076]).

Pour mémoire et à titre d'exemple, le logiciel d'aide à la détection de fracture a permis de corriger 14 examens radiologiques faussement interprété par l'Interne, en première lecture, comme « normaux ». Cet accroissement de la sensibilité de détection des fractures est important car les complications et la morbidité d'une fracture passée inaperçue peuvent être significatives.
De surcroit, avec l'aide de l'IA, l'Interne a diagnostiqué 96% des fractures du poignet, 94% des fractures du coudes, 100% des fractures du genou et 92% des fractures de la cheville. Sans l'aide de l'IA, ces pourcentages n'étaient que de 78% des fractures du poignet, 67% des fractures du coudes, 88% des fractures du genou et 76% des fractures de la cheville.

Peu d'études ont, à notre connaissance, évalué les performances et concordances diagnostiques entre différentes interprétations formulées par des professionnels de la santé assistés par l'IA, dans la recherche de fractures osseuses pédiatriques du squelette

appendiculaire. En outre, à contrario de notre étude portant sur de multiples articulations, les quelques études publiées chez l'enfant, examinant les performances diagnostiques d'un logiciel de *deep-learning* pour la détection de fractures osseuses, se sont concentrées sur des articulations précises telles que celle du coude (18)(19,20), du poignet (2)(24)(25) ou encore de la hanche (26)(27). Les études d'England et al. et de Choi et al. sont encore plus spécifiques, se focalisant respectivement sur les fractures supra-condyliennes (19) et les épanchements articulaires du coude (20). Cette restriction du champ d'étude à une articulation unique ou à un type de fracture réduit nettement le potentiel de transposition à la pratique radiologique courante.

Bien que focalisées sur certaines articulations, lesdites études ont cependant toutes montré une amélioration significative des performances diagnostiques des praticiens avec le concours du logiciel de *deep-learning*. Cette amélioration des performances était particulièrement marquée chez des médecins peu expérimentés (Interne en début de cursus) (2), non spécialisés (médecin urgentiste, chirurgien, etc.) ou encore lors d'un travail nocturne. Les logiciels d'intelligence artificielle ont, par le biais de ces études, révélé des performances diagnostiques remarquables non influencées par des facteurs extérieurs tels que la fatigue, le stress ou la charge de travail (24).

D'autre part, quelques études publiées, chez l'adulte, examinant les logiciels d'IA pour la détection de fractures osseuses, utilisent des *viewers* (2) distincts des PACS pour l'interprétation des imageries. L'usage de deux systèmes d'interprétation est pourtant peu confortable en pratique courante. De plus, certaines études, comme l'étude pédiatrique de Rayan et al. (18), font usage d'algorithmes nécessitant un formatage des données d'imagerie avec une perte d'information (500 x 500 pixels) limitant fortement la capacité diagnostique des lésions discrètes ou de petites tailles. Alors que la plupart des études précédemment citées utilisent des images recadrées ou réduites, le système d'IA utilisé dans notre étude traite des images en pleine résolution avec plusieurs radiographies par patient et peut être directement intégré dans le *Work Flow* et le PACS[3].

[3] Le *Picture Archiving and Communication System* (PACS) est le système de gestion électronique des images médicales avec des fonctions d'archivage, de stockage et de communication rapide. Il est utilisé en complément du système réseautique de gestion des activités des services de radiologies, appelé Système d'Information Radiologique (SIR), pour la gestion des images. Comme le souligne la Société Française de Radiologie (SFR), c'est « un outil clé pour la prise en charge des patients et la cohérence des soins » (28).

Pour rester au plus près de la réalité du terrain, a également été conservé, l'ensemble des examens radiographiques de mauvaise qualité ainsi que les diagnostics évidents ou présentant une inhabituelle difficulté.

Il est à noter que la conception de l'étude a permis à tous les « lecteurs » d'interpréter la même imagerie avec et sans IA. Cette conception favorise une plus grande puissance statistique des résultats.

D'autre part, pour éviter un biais contextuel et garder une ligne de conduite proche de la pratique quotidienne des praticiens, les radiologues avaient accès, lors de l'interprétation des imageries, au « bon d'examen » rédigé par l'urgentiste ainsi qu'aux données cliniques du patient. Cette ligne de conduite n'a pas été, à notre connaissance, utilisée dans d'autres études. Lindsey et al. (2) déplorent d'ailleurs dans leur discussion un manque de transmission d'informations concernant l'évaluation globale du patient par le clinicien.

Contrairement à certains articles publiés sur la détection de fracture à l'aide d'un logiciel d'IA(7), la présente étude a considéré l'évaluation de la détection des fractures comme une tâche de classification binaire, excluant l'identification de fractures multiples sur une seule image. Cette méthodologie d'interprétation semble correspondre davantage à la pratique quotidienne. Une conclusion dichotome évoquant la « présence » ou « l'absence » d'une fracture influence majoritairement la prise en charge. La topographie, le nombre ou encore le type de fracture modifient plus rarement la prise en charge du patient.

Bien que l'étude présente une conception proche de la réalité du terrain, la dichotomie des conclusions peut être source d'erreurs d'interprétation. Après avoir identifié une première anomalie, les radiologues observés auront tendance à cesser de continuer à rechercher d'autres lésions. Ce biais d'interprétation est connu sous le nom de « satisfaction de la recherche » (29).

Il doit également être notifié, qu'afin d'être en adéquation avec les capacités de détection du logiciel Boneview, que la recherche d'épanchement sur les imageries recueillies a été exclue. Un tel épanchement ne pouvant être détecté par l'IA que dans la région du coude, sa prise en compte aurait entravé l'homogénéité des résultats. Ce biais est important car la visualisation

d'un épanchement est souvent le seul signe de fracture du coude chez l'enfant en raison des fractures incomplètes ou cartilagineuses.

Outre ce biais potentiel d'interprétation, comme beaucoup d'autres études pédiatriques publiées (18,20), l'étude fait usage exclusif d'échantillon monocentrique et rétrospectif. De surcroît, le contexte d'interprétation en urgence associé, le cas échéant, à des horaires de nuit, n'a pu être reproduit en raison du caractère rétrospectif de l'étude. Une étude prospective multicentrique incluant l'interprétation d'un médecin urgentiste permettrait d'évaluer l'apport direct d'un logiciel de *deep-learning* au sein des services d'accueil des urgences.

Bien qu'en pratique courante les interprétations réalisées par un Professeur de radiologie constituent le *gold standard*, l'adjonction d'une seconde lecture « senior » associée à une consultation clinique de suivi ou à des radiographies de contrôle, permettrait la consolidation du diagnostic. Un diagnostic, obtenu de la sorte, serait indiscutable.
Il est à noter, qu'en l'absence d'un consensus quant à la manière d'affirmer avec certitude la présence d'une lésion, la définition d'un *gold standard* porte à caution. Dans la majorité des études précitées le *gold standard* est constitué d'une simple ou double lecture par un radiologue expérimenté dans la surspécialité étudiée. Le choix d'un tel référentiel, bien que discutable, est en accord avec la pratique clinique courante qui prend pour référence l'expérience.

Par ailleurs, la conception de l'étude laisse place à deux effets d'interprétation l'un concerne uniquement l'Interne et l'autre l'ensemble des lecteurs.

Par effet d'entraînement, l'Interne a vu au fil des 400 interprétations une incontestable amélioration de ses performances d'identification des lésions traumatiques et, par conséquent, de sa concordance diagnostique avec le senior. L'expérience est indubitablement source d'acquisition de compétences.

L'effet Hawthorne a également pu affecter les lecteurs, conduisant, à une modification de leur comportement en réponse à leur conscience d'être observés. Une lecture plus approfondie, qu'à l'accoutumée, des examens radiographiques, influence directement les performances diagnostiques.

En outre, un défaut de confiance dans les performances de l'intelligence artificielle a conduit l'Interne à modifier son compte-rendu dans le seul cas où, le logiciel Boneview avait décelé une fracture non visualisée lors de sa première interprétation. A contrario, les comptes rendus n'ont fait l'objet d'aucune normalisation dès lors que l'IA signifiait une absence de fracture. La prise en considération de l'aide-diagnostique du logiciel dans certains de ces dossiers aurait amélioré les performances-diagnostiques. Une meilleure connaissance de la spécificité et des valeurs prédictives positives reliées aux caractéristiques intrinsèques de détection des fractures du logiciel, permettrait d'abolir les préjugés des praticiens.

Les résultats obtenus soulèvent parallèlement d'autres réflexions concernant, d'une part, la disparité d'examens pathologiques selon l'articulation étudiée et, d'autre part, la performance de détection des fractures pédiatriques du logiciel Boneview®.

Après l'interprétation par un senior des imageries recueillies, l'analyse des résultats montre, en fonction de l'articulation explorée, une forte disparité des taux d'examens présentant une anomalie dite pathologique. Seuls 9 % des examens intéressants le genou sont pathologiques contre 25 % concernant les chevilles, 36 % concernant les coudes et 54 % concernant les poignets. Cette disparité prête à se questionner sur la pertinence des indications d'imagerie lors de l'exploration des traumatismes pédiatriques, notamment du genou. Une rationalisation de l'usage de la radiographie pourrait abaisser le nombre d'explorations irradiantes de l'enfant.

D'autre part, à l'inverse de ce à quoi on pouvait s'attendre, au vu des constations faites dans certaines études de performances utilisant le logiciel Boneview® chez l'adulte (7), la présente étude révèle que l'Interne de radiologie, non assisté par l'IA, présente de meilleures performances et concordances diagnostiques avec le senior, dans la détection des fractures du squelette appendiculaire de l'enfant, que le logiciel seul. Ces différences sont cependant non statistiquement significatives.
Les différentes phases de croissance osseuse et les variantes anatomiques de l'enfant peuvent prêter à confusion et pourraient se révéler être une source d'erreurs d'interprétation pour ledit logiciel. Face à la complexité diagnostique des fractures pédiatriques, les logiciels d'IA devraient être développés spécifiquement pour des radiographies d'enfants avec des kits d'apprentissage adaptés. Leurs performances de détection devraient être éprouvées au cours d'études prospectives multicentriques dédiées.

Enfin, après une analyse approfondie des fractures décelées en excès ou à défaut par l'IA, il semble que celle-ci soit performante dans la détection des interruptions et déformations corticales. Cette sensibilité de détection des anomalies corticales permet dans certains cas de récupérer de petites lésions non visualisées par le radiologue (**Figure 6**). Cependant, l'IA semble rencontrer des difficultés à déceler les fractures sans déformations corticales tels que les avulsions ou arrachements ostéo-chondraux (**Figure *8***). De plus, certaines imageries faussement positives peuvent être des superpositions d'images, créant une pseudo-déformation corticale, ou des lésions corticales non traumatiques (**Figure *7***), détectées à tort par le logiciel Boneview® comme une fracture. Il existe en conséquence une piste d'amélioration pour permettre au logiciel d'IA de visualiser plus de sous-types de fractures et de ne pas positiver toute anomalie corticale.

CONCLUSION

L'utilisation d'un logiciel d'IA permet à un Interne de radiologie, en fin de cursus, d'améliorer ses performances diagnostiques, avec comme référence un Professeur de radiopédiatrie expérimenté, concernant la détection de fractures pédiatriques appendiculaires grâce à une amélioration significative de sa sensibilité de détection.
En pratique, moins de fractures pédiatriques des membres passeraient inaperçues lors d'un passage aux urgences. Néanmoins, il reste nécessaire d'améliorer les performances diagnostiques pédiatriques de l'IA qui se montre moins efficiente que chez l'adulte (7).

La singularité de la présente étude réside dans le fait d'étudier tous types de lésions traumatiques, sur les 4 principales articulations de l'enfant, tout en étant proche d'une activité « en vie réelle » avec l'utilisation du bon d'examen et d'un logiciel d'IA directement intégré au PACS du centre hospitalier.

L'amélioration et la généralisation de cet outil, permettraient aux radiologues non spécialisés en radiologie-pédiatrique, de réaliser une interprétation des radiographies traumatiques du squelette appendiculaire plus efficiente en réduisant au maximum les radiographies faussement négatives. L'utilisation de logiciels d'intelligence artificielle pourrait ainsi pallier partiellement à la faible démographie en radiologues pédiatriques dans de nombreux centres hospitaliers en réduisant le nombre d'erreurs diagnostiques et toutes les conséquences juridiques et financières qui en résultent. La qualité et la sécurité des soins fournis dans les centres sous dotés en spécialistes, s'en trouveraient améliorées.

En outre, l'utilisation de logiciels d'IA pourrait réduire l'épuisement professionnel des radiologues (30) par l'amélioration des conditions de travail. Les coûts de santé, au fil des ans, s'en trouveraient également abaissés par une diminution des poursuites judiciaires et des dédommagements et l'amélioration de la prise en charge des populations pédiatriques.

Néanmoins, le coût, difficilement supportable pour certaines structures hospitalières lié à l'acquisition de ces logiciels reste une limite à leur emploi. En outre, bien que le logiciel Boneview® soit détenteur d'un marquage CE sur le territoire européen, la certification de tels logiciels novateurs, à titre de dispositifs médicaux, reste délicate dans certaines régions du monde (18)(31).

BIBLIOGRAPHIE

1. Poitou P, Loge I, Hastier-Gouin N, Guyet S, Belgaid A, Dufour D, et al. P508 - Motivation des consultations aux urgences pédiatriques. Arch Pédiatrie. 1 juin 2010;17(6, Supplement 1):177.
2. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 6 nov 2018;115(45):11591-6.
3. Arasu VA, Abujudeh HH, Biddinger PD, Noble VE, Halpern EF, Thrall JH, et al. Diagnostic emergency imaging utilization at an academic trauma center from 1996 to 2012. J Am Coll Radiol JACR. mai 2015;12(5):467-74.
4. Care Quality Commission. A national review of radiology reporting within the NHS in England.
5. W. Anderson M, Keats T. Atlas of Normal Roentgen Variants That May Simulate Disease - 9th Edition. 2012.
6. DeFroda SF, Hansen H, Gil JA, Hawari AH, Cruz AI. Radiographic Evaluation of Common Pediatric Elbow Injuries. Orthop Rev. 20 févr 2017;9(1):7030.
7. Duron L, Ducarouge A, Gillibert A, Lainé J, Allouche C, Cherel N, 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. juill 2021;300(1):120-9.
8. M. Daniel CHASSEING. Rapport d'information de la Cour des comptes sur l'adaptation aux besoins des moyens matériels et humains consacrés à l'imagerie médicale, 2016. Report No.: 602.
9. Tranovich MJ, Gooch CM, Dougherty JM. Radiograph Interpretation Discrepancies in a Community Hospital Emergency Department. West J Emerg Med. 1 juill 2019;20(4):626-32.
10. Pennsylvania Patient Safety Advisory. Communication of radiograph discrepancies between radiology and emergency departments.
11. Guly H. Diagnostic errors in an accident and emergency department. Emerg Med J EMJ. juill 2001;18(4):263-9.
12. Teixeira PGR, Inaba K, Salim A, Rhee P, Brown C, Browder T, et al. Preventable morbidity at a mature trauma center. Arch Surg Chic Ill 1960. juin 2009;144(6):536-41; discussion 541-542.

13. Whang JS, Baker SR, Patel R, Luk L, Castro A. The causes of medical malpractice suits against radiologists in the United States. Radiology. févr 2013;266(2):548-54.
14. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med. 5 avr 2007;356(14):1399-409.
15. Shaukat F, Raja G, Frangi AF. Computer-aided detection of lung nodules: a review. 2019;
16. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. févr 2017;542(7639):115-8.
17. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. janv 2020;577(7788):89-94.
18. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A. Binomial Classification of Pediatric Elbow Fractures Using a Deep Learning Multiview Approach Emulating Radiologist Decision Making. Radiol Artif Intell. 30 janv 2019;1(1):e180015.
19. England JR, Gross JS, White EA, Patel DB, England JT, Cheng PM. Detection of Traumatic Pediatric Elbow Joint Effusion Using a Deep Convolutional Neural Network. AJR Am J Roentgenol. déc 2018;211(6):1361-8.
20. Choi JW, Cho YJ, Lee S, Lee J, Lee S, Choi YH, et al. Using a Dual-Input Convolutional Neural Network for Automated Detection of Pediatric Supracondylar Fracture on Conventional Radiography. Invest Radiol. févr 2020;55(2):101-10.
21. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology. avr 2018;287(1):313-22.
22. Zhang-Yin JT. Évaluation de la concordance dans les études d'imagerie diagnostique : une étude de la qualité des données publiées. 18 juin 2015;69.
23. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977;33(1):159-74.
24. Blüthgen C, Becker AS, Vittoria de Martini I, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. Eur J Radiol. mai 2020;126:108925.
25. Yahalomi E, Chernofsky M, Werman M. Detection of distal radius fractures trained by a small set of X-ray images and Faster R-CNN. ArXiv181209025 Cs, 21 déc 2018.

26. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. ArXiv171106504 Cs Stat, 17 nov 2017.
27. Cheng C-T, Ho T-Y, Lee T-Y, Chang C-C, Chou C-C, Chen C-C, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol. oct 2019;29(10):5469-77.
28. Les radiologues plaident pour le développement des Pacs. Disponible sur : https://www.ticsante.com/story/643/les-radiologues-plaident-pour-le-developpement-des-pacs.html
29. Brady AP. Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging. févr 2017;8(1):171-82.
30. Jalal S, Parker W, Ferguson D, Nicolaou S. Exploring the Role of Artificial Intelligence in an Emergency and Trauma Radiology Department. Can Assoc Radiol J J Assoc Can Radiol. févr 2021;72(1):167-74.
31. Chanet Marie, Les dispositifs médicaux : comparaisons et perspectives du droit européen et nord-américain, Bordeaux, LEH Édition, 2019, Mémoires numériques de la BNDS.

SERMENT D'HIPPOCRATE

Au moment d'être admis(e) à exercer la médecine, je promets et je jure d'être fidèle aux lois de l'honneur et de la probité.

Mon premier souci sera de rétablir, de préserver ou de promouvoir la santé dans tous ses éléments, physiques et mentaux, individuels et sociaux.

Je respecterai toutes les personnes, leur autonomie et leur volonté, sans aucune discrimination selon leur état ou leurs convictions. J'interviendrai pour les protéger si elles sont affaiblies, vulnérables ou menacées dans leur intégrité ou leur dignité. Même sous la contrainte, je ne ferai pas usage de mes connaissances contre les lois de l'humanité.

J'informerai les patients des décisions envisagées, de leurs raisons et de leurs conséquences.

Je ne tromperai jamais leur confiance et n'exploiterai pas le pouvoir hérité des circonstances pour forcer les consciences.

Je donnerai mes soins à l'indigent et à quiconque me les demandera. Je ne me laisserai pas influencer par la soif du gain ou la recherche de la gloire.

Admis(e) dans l'intimité des personnes, je tairai les secrets qui me seront confiés. Reçu(e) à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les moeurs.

Je ferai tout pour soulager les souffrances. Je ne prolongerai pas abusivement les agonies. Je ne provoquerai jamais la mort délibérément.

Je préserverai l'indépendance nécessaire à l'accomplissement de ma mission. Je n'entreprendrai rien qui dépasse mes compétences. Je les entretiendrai et les perfectionnerai pour assurer au mieux les services qui me seront demandés.

J'apporterai mon aide à mes confrères ainsi qu'à leurs familles dans l'adversité.

Que les hommes et mes confrères m'accordent leur estime si je suis fidèle à mes promesses ; que je sois déshonoré(e) et méprisé(e) si j'y manque.

ORDRE NATIONAL DES MÉDECINS
Conseil National de l'Ordre

Introduction - Les suspicions de fractures du squelette appendiculaire, représentent le premier motif de consultation au sein des urgences pédiatriques. Or, le territoire français présente une carence en radiologues avec une forte disparité territoriale d'autant plus marquée concernant la radiologie pédiatrique. Ce manque de praticiens spécialisés, se traduit fréquemment par l'interprétation des radiographies pédiatriques par les médecins urgentistes sans avis immédiat radiologique. Cette organisation est source d'erreurs de diagnostic. L'objectif de cette étude comparative est de démontrer qu'il est possible de pallier partiellement la faible démographie de radiologues spécialisés en radiopédiatrie grâce à un radiologue généraliste aidé d'un logiciel d'intelligence artificielle.

Matériels et Méthodes - Un échantillon de quatre séries de 100 examens radiographiques pour traumatismes des principales articulations que sont le genou, la cheville, le coude et le poignet, issues de patients âgés de moins de 16 ans, reçus au service des urgences pédiatriques de l'Hôpital Nord de Marseille, a été recueilli. Cet échantillon d'examens, correspondant à 804 radiographies, associées aux informations cliniques disponibles sur le « bon d'examen », rédigé par le médecin urgentiste, a été interprété par un Professeur universitaire spécialisé en radiopédiatrie ayant plus de 20 ans d'expérience, représentant le *gold standard*, et un Interne de radiologie, en fin de cursus. Par la suite, les imageries médicales ont été extraites du serveur de l'Assistance Public Hôpitaux de Marseille, anonymisées et envoyées, de manière sécurisée, à la société Gleamer. Après traitement de ces examens, par le logiciel d'aide à la détection de fracture Boneview, l'Interne a réalisé une nouvelle interprétation, en aveugle de sa précédente analyse. Les comptes rendus résultants desdites interprétations, rédigés par l'Interne et les Professeurs universitaires de radiopédiatrie, ont été classés de manière binaire en deux catégories : « présence d'une fracture ou doute sur l'existence d'une fracture » versus « pas de fracture visible sur l'imagerie ». Suite à ces interprétations un tableau de contingences a été construit, permettant d'extraire les données de performance diagnostique. La concordance diagnostique a également été évaluée grâce au calcul du coefficient Kappa de Cohen. La significativité des résultats a ensuite été mesurée grâce à un test de Student sur données appariées.

Résultats - Le concours du logiciel d'aide à la détection de fracture a amélioré la performance diagnostique de l'Interne. Sa sensibilité de détection s'est significativement accrue passant de 0.82 à 0.94 (différence absolue : 0.12 ; intervalle de confiance à 95% [+0.1271 ; +0.1076]). La valeur prédictive négative est également, de manière non significative, améliorée. Toutefois, le pendant, non significatif, de cette amélioration est une baisse de la spécificité et de la valeur prédictive psitive. Plus concrètement, l'aide de l'IA a permis, sur 400 examens radiologiques, de détecter 14 dossiers faussement classés en « absence de fracture » par l'Interne, lors de sa première interprétation. En outre, sans significativité statistique, la concordance diagnostique de l'Interne avec le Professeur de radiopédiatrie, s'est également accrue passant de 0.76, sans l'aide l'IA, à 0.83 (différence absolue : 0.06 ; intervalle de confiance à 95% [0.1548 ; -0.031]).

Conclusion - La présente étude démontre que l'utilisation d'un logiciel d'intelligence artificielle permet à un Interne de radiologie, en fin de cursus, d'améliorer ses performances diagnostiques avec un Professeur de radiopédiatrie expérimenté concernant la détection de fractures sur des radiographies traumatiques pédiatriques avec une amélioration significative de la sensibilité de détection. La généralisation de cet outil permettrait aux radiologues, non spécialisés en pédiatrique, de réaliser des interprétations plus efficientes. La qualité et la sécurité des soins fournis dans les centres sous-dotés en spécialistes, s'en trouveraient améliorées.

Mots clés - Intelligence artificielle, apprentissage profond, radiologie d'urgence et de traumatologie, radiologie pédiatrique, radiographie.