Validation externe des performances diagnostiques d'un logiciel d'intelligence artificielle pour le diagnostic radiologique des fractures du coude de l'enfant

Julie da Costa

Service Commun de la Documentation
Chemin la censive du Tertre - BP 32211
44322 Nantes cedex 3

www.bu.univ-nantes.fr

NANTES UNIVERSITÉ

_____

FACULTÉ DE MEDECINE

_____

Année 2022

N°

THESE

pour le

DIPLOME D'ETAT DE DOCTEUR EN MEDECINE

(DES de MÉDECINE GÉNÉRALE)

par

Julie DA COSTA

_____

Présentée et soutenue publiquement le 18 octobre 2022

_____

VALIDATION EXTERNE DES PERFORMANCES DIAGNOSTIQUES D'UN LOGICIEL D'INTELLIGENCE ARTIFICIELLE POUR LE DIAGNOSTIC RADIOLOGIQUE DES FRACTURES DU COUDE DE L'ENFANT

_____

Présidente : Madame la Professeure GRAS-LE GUEN

Directrices de thèse : Docteure LORTON et Docteure VRIGNAUD

Sommaire

Remerciements

A Madame la Professeure Christèle Gras-Le Guen, vous êtes à l'origine de ce projet et me faites l'honneur de présider le jury de ma thèse. Je vous suis reconnaissante de l'attention soutenue que vous avez portée à ce travail ainsi que pour la chance que vous m'offrez d'approfondir mes connaissances en pédiatrie.

A Madame la Professeure Elise Launay, vous étiez présente lors de ma première garde dans votre service ainsi qu'à ma soutenance de thèse. Votre bienveillance a su me donner confiance en moi et j'espère que nous travaillerons encore ensemble.

A Monsieur le Professeur Frampas et Monsieur le Docteur Decante, la rapidité de votre aide dans ce travail a permis la réalisation de cette thèse. Votre réactivité et votre gentillesse ont été d'un précieux secours lors des périodes de doute.

A Mesdames les Docteures Bénédicte Vrignaud et Fleur Lorton, vous m'avez accompagnée avant même la naissance de ce projet puis en tant que directrices de thèse. Votre implication de tous les instants, dont je ne saurais assez vous remercier, est pour beaucoup dans la rapidité de réalisation de ce travail. Ce sera un plaisir pour moi de m'enrichir à vos côtés à l'avenir.

A Madame la Docteure Laura Meurice, ton travail en amont a été d'une aide précieuse. Nos travaux parallèles s'étayaient l'un l'autre et j'ai pris plaisir à échanger avec toi à leurs sujets. J'ai hâte de continuer à te côtoyer régulièrement.

A mes relecteurs, qui ont su polir cette thèse jusqu'à la moindre coquille en un temps record.

Remerciements personnels

A ma famille, Maman, Papa, Emilie et Victor, Louis et Ludivine, merci pour tout. Vous avez toujours été là pour moi, bien avant cette thèse éclair, bien avant l'ECN et son lot de stress, bien avant la PACES où vous avez été un soutien permanent. Vous m'avez offert un terreau fertile et soutenant pour grandir et faire mes choix. En cas de difficulté, vous réunissez le conseil de famille, quel que soit le sujet, et dépensez votre temps sans compter. La distance ne nous a jamais séparés ni amoindri l'amour que je vous porte, et c'est avec confiance que j'envisage l'avenir avec nos retrouvailles comme points de repère.

A Ophélie, tu occupes une telle place dans ma vie qu'elle serait trop calme sans toi. Merci pour ton soutien, tes diversions, ton oreille attentive. Tu sais nourrir mon élan et faire diversifier mes projets avec une débauche d'imagination délicieuse.

A Mathilde, tu es la fille parfaite. Tu m'as accompagnée depuis le début de cette grande aventure de la médecine et dans bien d'autres domaines. A toute heure du jour et de la nuit je peux trouver auprès de toi des conseils, du réconfort et des smoothies.

A ma sous-colle, vous êtes la preuve que l'adversité forge les plus belles amitiés. A toi, Théo, mon jumeau de thèse, j'apprécie tout particulièrement ta curiosité insatiable qui te pousse hors des chemins communs (et des recettes françaises). A toi, Séverin, tu me surprends à chaque détour de conversation par l'acuité de ton jugement et la diversité de tes centres d'intérêts. A toi, Adrien, ton humour sait éclairer mes journées. Tous les trois, vous me faites découvrir chaque jour des possibilités dont je n'aurais jamais supposé l'existence et qui stimulent mon ouverture d'esprit.

A mes amis de lycée, le Keur, merci de grandir avec moi depuis plus de dix ans. Nos discussions, nos sorties, nos centres d'intérêt continuent d'évoluer au fil du temps, d'étendre des ramifications dans toutes les directions depuis le tronc commun de notre amitié au fur et à mesure que nous avançons dans la vie. J'adore passer de branche en branche pour profiter des personnalités et des originalités de chacun, tout comme j'adore nos réunions qui rassemblent ces ramures pour créer de nouveaux souvenirs communs.

A mes amis d'externat, les Pamplemousses, je chérie votre originalité plus que tout. Bien que nos internats respectifs nous aient éparpillés aux quatre coins de la France, je vous retrouve toujours avec plaisir.

A mes amis d'internat, vous avez su peupler Nantes qui était une ville inconnue pour moi. A Léna, ton humour et ta passion m'inspirent. A Justine, Elise et Anne-Sophie, l'internat était beaucoup moins vide grâce à vous. Aux dix-sept co-internes de pédiatrie, notre ambiance de travail et de sorties sont l'un de mes meilleurs souvenirs de stage. A Aurélie, tu incarnes pour moi l'essence de l'urgentiste et j'admire ta force de caractère. A l'équipe de Saint-Nazaire, Romain, Loïc, Elodie, Philippe, Morgane, PA, Camille, Cécilia et Louis, nos sorties régulières sont de vraies bouffées d'oxygène et cela s'est confirmé lors de cette semaine dans le Vercors où vous m'avez soutenue dans les vires et revires, tant montagnardes qu'analytiques.

A Emma et Amandine, vous avez assisté du début à la fin à la réalisation de cette thèse dans les moindres détails. Merci pour votre soutien et vos coups de pouce précieux.

A ma fanfare, les Makabés, vous êtes l'un des virages les plus improbables et les plus absolus qui émaillent mon chemin. Je me suis découvert à vos côtés, non pas un talent pour la musique, mais d'autres aspects étonnants de ma personnalité. Vous m'enrichissez à chacun de mes passages parmi vous malgré la différence d'âge qui se creuse.

Ces trop courts remerciements ne rendent pas honneur à toutes les personnes merveilleuses qui ont croisé ma vie. J'ai encore tant d'autres noms que je ne peux ajouter ici, mais soyez sûrs que je pense également à vous qui n'avez pas été cités. A tous donc, merci !

Abréviations

BABP	Plâtre brachio-anté-brachio-palmaire
CHU	Centre hospitalo-universitaire
CLAIM	Checklist for Artificial Intelligence in Medical Imaging
DICOM	Digital imaging and communications in medicine
EIQ	Ecart interquartile
Fig.	Figure
IA	Intelligence artificielle
$IC_{95\%}$	Intervalle de confiance à 95%
PACS	Picture archiving and communication system
RV+	Rapport de vraisemblance positif
RV-	Rapport de vraisemblance négatif
Se	Sensibilité
Sp	Spécificité
VPN	Valeur prédictive négative
VPP	Valeur prédictive positive

Sommaire des tableaux et des figures

Introduction

Les fractures du coude sont fréquentes dans la population pédiatrique et représentent en moyenne 15 à 20% de l'ensemble des fractures de l'enfant (1). Cependant, les radiographies de cette articulation sont difficiles à analyser et la fréquence des fractures initialement non diagnostiquées varie de 17 à 77% dans les études pédiatriques (2,3) contre 6% chez les adultes (4). Cette complexité s'explique par l'apparition successive des noyaux d'ossification (condyle externe à 2 ans, tête radiale à 4 ans, épicondyle médial à 6 ans, trochlée à 8 ans, olécrane à 10 ans, et épicondyle latéral à 12 ans (1)), nécessitant la recherche d'indicateurs indirects de fracture comme l'hémarthrose (5). Les enjeux diagnostique et thérapeutique sont importants car le remodelage osseux y est faible. Ce phénomène implique de multiples complications chez les fractures négligées : 58% des fractures supra condyliennes mal traitées se compliquent de cubitus varus, qui comme les pseudarthroses du coude peuvent entrainer des lésions tardives du nerf ulnaire, des douleurs et une instabilité de l'articulation (6). Par ailleurs, la demande croissante de radiographies pour motifs traumatologiques, plus particulièrement dans les services d'urgences, ne peut permettre une lecture rapide par un spécialiste. La première analyse de ces radiographies revient aux internes et urgentistes (7) qui ne sont pas forcément des experts dans ce domaine. Les fractures passées inaperçues représentent jusqu'à 80% des erreurs diagnostiques des services d'urgences (8). Enfin, un doute radiologique peut amener à de nouveaux clichés impliquant une irradiation plus importante et une augmentation des coûts (9).

Dans ce contexte, l'intelligence artificielle (IA) pourrait constituer une aide au diagnostic. Si les premiers essais n'ont pas été concluants (10), l'essor récent du *deep learning* a permis une nouvelle approche : plutôt que de donner une suite d'ordres spécifiques au logiciel, les développeurs se sont inspirés du fonctionnement du cerveau humain pour créer un système de neurones s'éduquant à partir de bases de données (11). L'augmentation des capacités informatiques autorisent maintenant des logiciels à plusieurs centaines de couches de neurones dialoguant entre elles, aptes à des interprétations complexes (9,12). Cette méthode offre un grand niveau d'abstraction qui permet de conserver l'apport de précédents entrainements malgré un changement de problème. Ainsi, un logiciel entrainé sur des images non médicales peut ensuite être validé pour la lecture de radiographies (10,12). Ce mécanisme a amélioré les performances des machines jusqu'à obtenir des résultats similaires aux radiologues ou aux orthopédistes, voire même surpasser les performances des professionnels (13,14). Duron a également montré que l'aide de l'intelligence artificielle permettait d'améliorer les performances de lecture des urgentistes sur les radiographies de membres en majorant leur sensibilité (Se) de 8,7% et leur spécificité (Sp) de 4,1% tout en réduisant le temps de lecture de 15% (7). C'est une aide reproductible et indépendante de facteurs extérieurs, alors que 40% des erreurs en service d'urgences sont liées à un défaut de vigilance ou de mémoire du praticien, 23% à la surcharge de travail (15), et qu'on observe un pic d'erreur de diagnostics radiographiques de 20% dans ces services entre 20h et 2h du matin (16).

Cependant, la plupart de ces études concerne les fractures chez l'adulte, dont les radiographies diffèrent de manière importante de la population pédiatrique en raison de l'absence de cartilage de croissance et de noyaux d'ossification. Des approches d'IA de type *deep learning* ont néanmoins récemment montré des performances diagnostiques prometteuses sur les radiographies pédiatriques. Dupuis a effectué une validation externe du logiciel Rayvolve sur 2 634 examens pédiatriques constitués de deux radiographies de membre montrant une Se de 95,7% (Intervalle de confiance à 95% ($IC_{95\%}$) 94,0-96,9) et une Spe de 91,2% ($IC_{95\%}$ 89,8-92,5) pour la détection des fractures (17). Hayashi a effectué une validation externe du logiciel Boneview sur 300 examens pédiatriques. Une analyse en sous-groupe concernant le coude sur 60 examens retrouvait une Se de 100,0% ($IC_{95\%}$: 88,4-100,0) et une Sp de 83,3% ($IC_{95\%}$ 73,5-97,9) ; néanmoins la proportion d'examens anormaux était de 50% (18). Concernant le coude pédiatrique, une revue systématique de la littérature (19) portant sur quatre études (17,20–22) pour un total de 2 156 examens a montré une Se comprise entre 88,9 et 90,7% et une Sp entre 90,9 et 100%. Cependant, seules deux études comportaient une validation externe (17,22) dont une pour laquelle l'étude du coude était une analyse en sous-groupe et non l'objectif principal (17).

L'objectif de cette thèse était ainsi d'effectuer une validation externe des performances diagnostiques du logiciel d'IA Boneview en vie réelle pour la détection des fractures du coude de l'enfant, articulation particulièrement complexe pour les cliniciens et dont les enjeux diagnostiques et thérapeutiques sont importants.

Méthodes

1. *Méthodologie générale et population d'étude*

Cette thèse est une étude observationnelle rétrospective de validation externe d'un logiciel d'IA pour le diagnostic de fracture du coude chez l'enfant. La validation externe consiste à tester le logiciel sur une population différente de celle sur laquelle il a été construit. Dans ce travail, la validation externe était à la fois géographique et temporelle car la population d'étude a été recrutée dans un autre hôpital et sur une autre période de temps que lors du développement des logiciels. Tous les enfants âgés de 0 à 15 ans et 3 mois admis aux urgences pédiatriques du Centre Hospitalier Universitaire (CHU) de Nantes à la suite d'un traumatisme du coude entre le 1er janvier 2019 et le 1er avril 2020 et pour lesquels une paire de radiographies du coude (face et profil) a été prescrite ont été inclus rétrospectivement et de façon exhaustive. Pour chaque patient, l'âge et le sexe ont été recueillis.

Les résultats ont été rapportés selon les recommandations CLAIM (Checklist for Artificial Intelligence in Medical Imaging)(23).

2. *Objectifs de l'étude*

L'objectif principal était l'évaluation des performances diagnostiques pour la détection des fractures du coude du logiciel d'IA Boneview développé par Gleamer. Les critères de jugement principal étaient la sensibilité (Se), la spécificité (Sp), les valeurs prédictives positive (VPP) et négative (VPN) ainsi que les rapports de vraisemblance positif (RV+) et négatif (RV-) du logiciel. Un examen était constitué d'un ensemble de deux radiographies du coude, face et profil. L'interprétation par le logiciel se faisait selon trois modalités de réponse : examen négatif, positif ou doute. L'examen était classé en « normal » si les deux incidences face et profil présentaient la réponse « négatif », et en « anormal » si au moins l'une des incidences présentait la réponse « positif » ou « doute ».

Les objectifs secondaires étaient :

- l'évaluation des performances diagnostiques du logiciel dans le cas où les réponses « doute » étaient classées avec les examens normaux

- la description des patients (âge, sex-ratio et type de fractures) dont les examens étaient classés incorrectement par le logiciel (faux-positifs et faux-négatifs) et ceux dont les examens étaient classés en « doute »

- l'évaluation des performances diagnostiques du logiciel en excluant des analyses les patients dont l'interprétation radiographique par le logiciel était classée en « doute »

3. *Ground truth ou vérité terrain*

Pour constituer la référence appelée « ground truth » ou « vérité terrain », les examens ont été relus indépendamment et à l'aveugle de l'histoire clinique, du diagnostic initial du clinicien et des résultats du logiciel par deux experts : un radiologue (parmi trois radiologues de 3 à 30 années d'expériences) et un pédiatre urgentiste spécialisé en traumatologie (parmi trois pédiatres de 4 à 16 années d'expérience). En cas de désaccord entre les deux premiers experts, l'avis d'un orthopédiste infantile avec 11 années d'expérience a été demandé. L'accord entre experts a été mesuré par le coefficient Kappa de Cohen (24) et interprété selon l'échelle de Landis et Koch (25).

La lecture des images radiologiques s'est effectuée au format DICOM (Digital Imaging and communications in medicine) sur le PACS (picture archiving and communication system), logiciel utilisé habituellement en pratique clinique (Carestream Vue PACS, version 11). Cette interprétation de référence était binaire : examen normal ou anormal. Les examens anormaux comprenaient les hémarthroses isolées, les fractures et les luxations de coude. En cas d'examen anormal, celui-ci était classé selon les catégories suivantes : hémarthrose isolée, fracture supra condylienne, du condyle externe, de l'épicondyle médial, du col radial, de l'olécrâne, luxation de coude, ou autres. Les fractures supra condyliennes ont été divisées en stade 1 à 4 selon la classification de Lagrange et Rigault (26).

4. *Modèle d'IA évalué*

Le logiciel étudié est Boneview dans sa version 2.0.3.1. C'est un algorithme basé sur Detrecton2, une plateforme de détection d'objet en open-source développée par Facebook research AI puis adaptée par Gleamer. Elle est écrite à partir de PyTorch, une base d'apprentissage contenant une bibliothèque pour les programmes Python qui facilite la construction des projets de *deep learning*. Le logiciel analyse les images au format DICOM et détermine les zones d'intérêt selon des normes de Se et de Sp déterminées à l'avance. Ces dernières ont été choisies lors du développement et de la validation interne du logiciel pour obtenir une valeur prédictive négative de 99,5%. La résolution des images n'a pas été modifiée et les analyses ont été faites au format DICOM. L'interprétation du logiciel Boneview pouvait donner trois types de réponse : examen négatif, positif ou doute. La réponse « doute » était donnée lorsque le seuil de confiance était compris entre 50 et 90% et la réponse « négatif » s'il était inférieur à 50%.

5. *Analyses statistiques et méthodes d'évaluation du modèle d'IA*

Premièrement, nous avons décrit les caractéristiques de la population, la fréquence et les types de fractures en présentant les effectifs et proportions ainsi que leur intervalle de confiance à 95% ($IC_{95\%}$) pour les variables qualitatives et la médiane et l'écart interquartile (EIQ) pour les variables quantitatives. Deuxièmement, les performances diagnostiques (Se, Sp, VVP, VPN, RV+, RV-) du logiciel ont été calculées à partir d'un tableau de contingence. Les probabilités pré- et post-test ont été représentées par un nomogramme de Fagan. Enfin, nous avons décrit les enfants dont les examens étaient classés

incorrectement ou en « doute » par le logiciel et conduit une analyse de sensibilité sur les performances diagnostiques en excluant les patients dont l'interprétation radiologique était classée en « doute ». En cas de données manquantes sur le résultat d'interprétation du logiciel, les patients étaient exclus de l'ensemble des analyses.

6. *Aspects réglementaires*

Il s'agissait d'une étude monocentrique, rétrospective sur données présentes dans le dossier médical du patient et considérée comme hors Loi Jardé. Les données des patients ont été recueillies en une fois dans une base de données anonymisée avec impossibilité de revenir sur l'identité du patient. L'information du patient et de ses représentants légaux à propos de l'utilisation de ses données de santé à visée de recherche s'est faite via le biais d'affiches d'information disposées dans le service des urgences pédiatriques.

Résultats

1. Description de la population

Durant la période d'étude, 767 enfants ont été inclus et 10 (1,3%) ont été exclus en raison d'une absence de données du logiciel (Fig. 1). Au total, les examens de 757 enfants ont pu être analysés. Le sex-ratio F/M était 1,0 (376 (49,7%) garçons, 381 (50,3%) filles) et l'âge médian de 8,3 ans (EIQ 5,2-11,7). La proportion d'examens anormaux était de 46,6% ($IC_{95\%}$ 43,1-50,2 ; n=353) comprenant 12,7% ($IC_{95\%}$ 10,3-15,1 ; n=96) d'hémarthroses isolées et 33,9% de fractures ou luxations ($IC_{95\%}$ 30,6-37,3 ; n=257). Les caractéristiques des patients et des radiographies sont décrites dans le Tableau 1.

Tableau 1 – Caractéristiques de la population d'étude

	Total des patients n=757
Sexe	
Fille	381 (50,3)
Garçon	376 (49,7)
Age	
Age médian, années (EIQ)	8,3 (5,2-11,7)
< 12 ans	590 (77,9)
Examens	
Normaux	404 (53,4)
Anormaux	353 (46,6)
- Hémarthroses isolées	96 (27,2)
- Fractures*	257 (72,8)
Supracondyliennes	125
Condyle externe	42
Epicondyle médial	19
Col radial	22
Olécrâne	21
Autres	39

Les données correspondent à l'effectif n=757 (100%)
**A noter que la somme des sous-catégories des fractures est supérieure au nombre de radiographies « fracture », car un même examen peut présenter plusieurs fractures.*

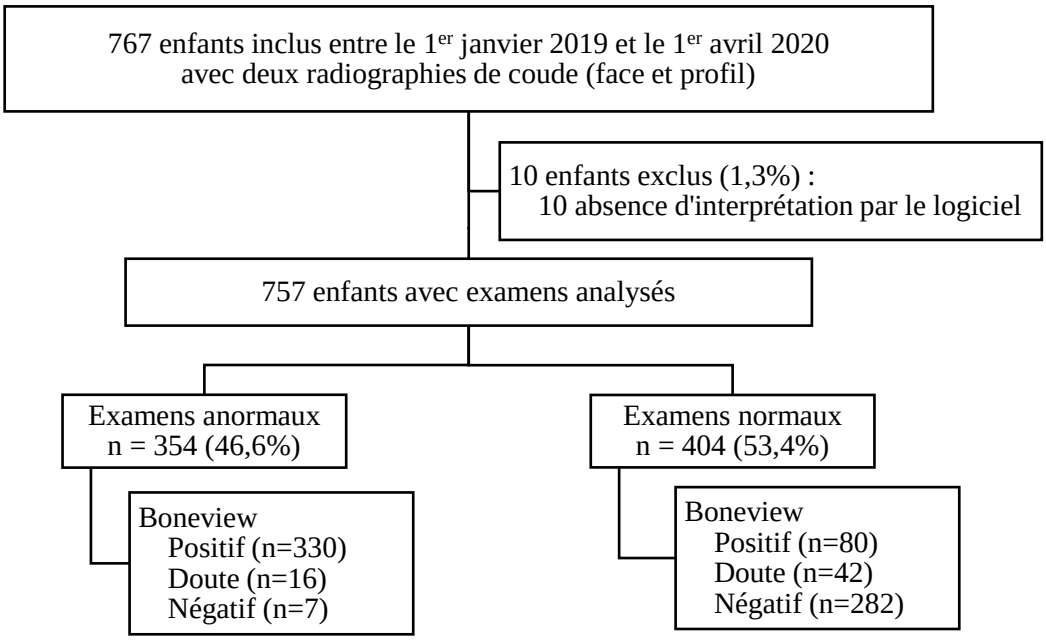

Fig. 1 – Diagramme de flux

2. *Accord entre les experts*

L'indice Kappa d'accord entre les deux premiers experts a été calculé à 0,77 ($IC_{95\%}$ 0,72-0,81), ce qui correspondait à un accord fort. Ils étaient en désaccord sur la classification de l'examen dans 11,6% des cas (n=88/757) (Tableau 2).

Tableau 2 – Désaccords entre les experts

	Radiologues	
	Examens anormaux	Examens normaux
Urgentistes		
Examens anormaux	300	16
Examens normaux	72	369

Les caractéristiques de l'effectif de ce sous-groupe sont décrites dans le Tableau 3. Parmi les 88 examens ayant entrainé un désaccord entre les deux premiers experts, 34 (38,6%) ont été classés comme normaux et 36 (40,9%) comme des hémarthroses isolées après relecture par le 3ème expert. Pour ces 88 examens, le 3ème expert était en accord avec l'interprétation du radiologue dans 58% (n=51/88) des cas et avec l'urgentiste dans 42% (n=37/88) des cas. Parmi les 34 examens classés comme normaux par le 3ème expert, 27 (79,4%) l'étaient aussi par l'urgentiste et 7 (20,6%) par le radiologue. Parmi les 54 examens classés comme anormaux par le 3ème expert, 10 (18,5%) l'étaient aussi par l'urgentiste et 44 (81,5%) par le radiologue. Les radiologues avaient tendance à produire des faux-positifs (n=27/37 soit 73,0%) tandis que les urgentistes avaient tendance à produire des faux-négatifs (n=44/51 soit 86,3%), notamment concernant les hémarthroses seules (n=32/51 soit 61,5%).

Tableau 3 – Caractéristiques des 88 patients pour lesquels un désaccord existait dans l'interprétation des deux premiers experts

	Patients n=88
Sexe	
Fille	30 (34,0)
Garçon	58 (66,0)
Age	
Age médian, années (EIQ)	9,5 (5,8-13,3)
< 12 ans	61 (69,3)
Examens*	
Normaux	34 (38,6)
Anormaux	54 (61,4)
- Hémarthroses isolées	36 (66,7)
- Fractures**	18 (33,3)
Supracondyliennes	4
Condyle externe	0
Epicondyle médial	4
Col radial	2
Olécrâne	4
Autres	5

Les données correspondent à l'effectif n=88 (11,6%)
**L'interprétation présentée est celle du 3ème expert.*
***A noter que la somme des sous-catégories des fracture est supérieure au nombre de radiographies « fracture », car un même examen peut présenter plusieurs fractures.*

3. Performances diagnostiques du logiciel

Parmi les 757 examens analysés, 54,2% ($IC_{95\%}$ 50,6-57,7 ; n= 410) étaient classés comme positifs, 38,2% ($IC_{95\%}$ 34,7-41,6 ; n= 289) comme négatifs et 7,7% ($IC_{95\%}$ 5,8-9,7 ; n= 58) en doute par le logiciel. Leur répartition selon le « ground truth » est exposée dans le Tableau 4.

Tableau 4 – Tableau de contingence

	Ground truth	
	Anormal (n=353)	Normal (n=404)
Boneview		
Positif (n=410)	330	80
Doute (n=58)	16	42
Négatif (n=289)	7	282

Ainsi, en regroupant les catégories « positif » et « doute », le logiciel avait une Se de 98,0% ($IC_{95\%}$ 96,0-99,2), une Sp de 69,8% ($IC_{95\%}$ 65,1-74,2), une VPP de 73,9% ($IC_{95\%}$ 71,0-76,7) et une VPN de 97,6% ($IC_{95\%}$ 95,1-98,8) (Tableau 5).

Tableau 5 – Performances diagnostiques du logiciel chez 757 enfants avec examen radiologique du coude

	Doutes classés en examens	
	Anormaux	Normaux
Se*	98,0% [96,0 – 99,2]	93,5% [90,4 – 95,8]
Sp*	69,8% [65,0 – 74,2]	80,2% [76,0 – 84,0]
VPP*	73,9% [71,0 – 76,7]	80,5% [77,2 – 83,4]
VPN*	97,6% [95,1 – 98,8]	93,37% [90,4 – 95,5]
RV+*	3,25 [2,80 – 3,77]	4,72 [3,87 – 5,76]
RV-*	0,03 [0,01 – 0,06]	0,08 [0,05 – 0,12]

**Les $IC_{95\%}$ sont indiqués entre crochets.*

Le RV+ était de 3,25 ($IC_{95\%}$ 2,80-3,77), ce qui correspond à un apport diagnostique modéré en cas de réponse positive du logiciel (27). Le RV- était de 0,03 ($IC_{95\%}$: 0,01-0,06), ce qui correspond à un apport diagnostique très fort en cas de réponse négative du logiciel (27). Ils sont représentés, ainsi que les probabilités pré et post-test, graphiquement par un nomogramme de Fagan (Fig. 2).

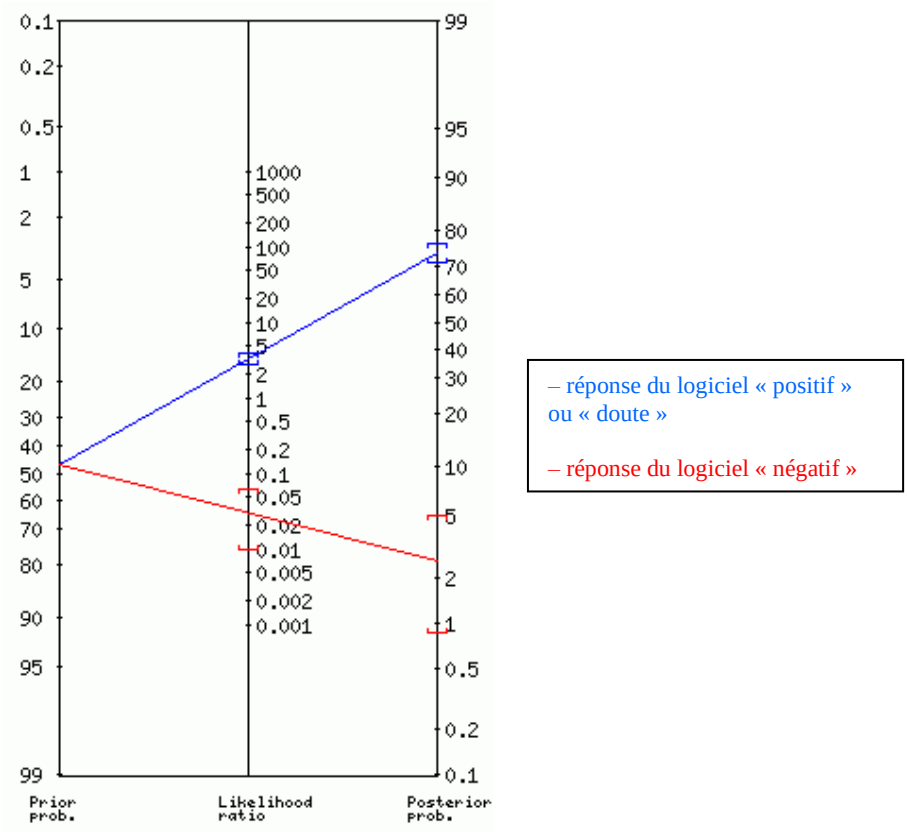

Fig. 2 – Nomogramme de Fagan

Les performances diagnostiques du logiciel en classant les examens interprétés comme « doute » dans la catégorie des examens normaux sont rapportées dans le Tableau 5.

4. *Description des patients avec un examen classé incorrectement ou en « doute » par le logiciel*

Chez les 80 enfants dont l'examen était classé à tort comme « positif » (faux-positifs), le sex-ratio F/M était de 1,3 (35 garçons et 45 filles), l'âge médian de 9,1 ans et 70% d'entre eux avaient moins de 12 ans. Chez les 7 enfants dont l'examen était classé à tort comme « négatif » (faux-négatifs), le sex-ratio F/M était de 0,4 (5 garçons et 2 filles), l'âge médian était de 2,0 ans et tous les enfants avaient moins de 5 ans. Dans ce cas, les anomalies radiologiques correspondaient à 5 hémarthroses isolées, une fracture supracondylienne de stade 1 et une fracture du condyle externe.

Chez les 58 enfants dont l'examen était classé en « doute », le sex-ratio F/M était de 0,8 (32 garçons et 26 filles), l'âge médian était de 8,8 ans et 86% avaient moins de 12 ans. Parmi ces 58 examens « doute », 42 (72,4%) étaient des examens normaux et 16 (27,6%) des examens anormaux selon le « ground truth » (). Les 16 examens anormaux se décomposaient en 3 hémarthroses isolées, 2 fractures supracondyliennes, 5 fractures du condyle externe, 3 fractures de l'épicondyle médial, 3 fractures de l'olécrâne et 2 luxations de coude.

5. *Analyse de sensibilité*

Les performances diagnostiques du logiciel calculées après exclusion des 58 patients avec un examen classé en « doute » sont rapportées dans le Tableau 6.

Tableau 6 – Performances diagnostiques du logiciel chez 699 enfants après exclusion des doutes

Se*	97,9% [95,8 – 99,2]
Sp*	77,9% [73,3 – 82,1]
VPP*	80,5% [77,3 – 83,4]
VPN*	97,6% [95,1 – 98,8]
RV+*	4,43 [3,65 – 5,38]
RV-*	0,03 [0,01 – 0,06]

**Les IC95% sont indiqués entre crochets.*

Discussion

Cette étude apporte une nouvelle validation externe d'un logiciel d'IA pour la lecture de radiographies alors que celles-ci sont encore rares dans la littérature, plus particulièrement dans la population pédiatrique (28,29). Le logiciel montrait ici une Se de 98,0% ($IC_{95\%}$ 96,0-99,2) et une Sp de 69,8% ($IC_{95\%}$ 65,1-74,2). Lorsque l'on compare ces résultats aux données des études portant sur le coude pédiatrique, la Se était plus élevée que celles rapportées dans la revue systématique de Shelmerdine (19), comprises entre 88,9% et 90,%. Cependant, cette étude concernait d'autres logiciels d'IA. La validation externe de Boneview menée par Hayashi retrouvait en revanche une Se élevée (Se 100,0% sur l'analyse en sous-groupe concernant le coude pédiatrique)(18). La Sp de cette étude était quant à elle inférieure aux chiffres de la littérature, compris entre 84 et 100% (17,19–22) ; cette tendance se retrouve également avec la validation externe de Boneview par Hayashi (Sp 90% sur l'analyse en sous-groupe concernant le coude pédiatrique)(18). Cette différence dans nos résultats sur la Sp pourrait s'expliquer par plusieurs facteurs. Premièrement, la qualité des radiographies ou le type de fracture ne constituaient pas des critères d'exclusion, contrairement à England (20) et Choi (22) ; les examens se rapprochaient donc de ceux effectivement disponibles en pratique clinique. Deuxièmement, le logiciel Boneview proposant la réponse « doute », il a été choisi de classer ces examens avec les examens anormaux pour nos analyses principales. En effet, le but de l'utilisation de ce logiciel en pratique courante étant avant tout d'exclure le diagnostic de fracture, la Se était privilégiée à la Sp. Or, les examens « doutes » correspondant finalement dans 72,4% des cas à des examens normaux, les classer avec les examens anormaux a fait diminuer la Sp. L'autre classification (« doutes » avec examens normaux) montrait en effet une meilleure Sp avec une différence significative, mais au détriment d'une diminution simultanée de la Se. Néanmoins, l'IA se destinant à améliorer les performance des lecteurs plutôt qu'à les remplacer (7), une grande Se permettrait d'attirer l'attention du lecteur sur les examens pouvant être anormaux afin que celui-ci statue sur la présence ou non d'une anomalie. Cette approche permet surtout de limiter les faux-négatifs, dont le nombre aurait triplé dans cette étude en classant les « doutes » avec les examens normaux plutôt qu'avec les examens anormaux (de 7 à 23). Cette démarche semble préférable au vu des complications importantes pouvant découler d'une fracture négligée (6).

Concernant les caractéristiques démographiques des enfants dont les examens étaient classés en « doute » ou qui étaient des faux-positifs, elles étaient assez proches de celles de l'ensemble de la population étudiée. Les faux-négatifs concernaient en revanche uniquement les patients les plus jeunes. Tous avaient moins de 5 ans, avec un âge médian de 2 ans contre 8 ans pour la population totale. Un entrainement du logiciel sur cette catégorie d'âge où l'articulation est majoritairement composée de cartilage avec peu de noyaux d'ossification constitue une piste possible d'amélioration de la Se du logiciel.

Cette étude a également mis en évidence l'existence de désaccords entre experts dans l'interprétation de radiographies pédiatriques dans 11,6% des cas, bien que le centre hospitalier participant soit un centre

spécialisé dans la traumatologie pédiatrique. 36 examens discordants sur 88, soit environ 40%, étaient des hémarthroses isolées, témoignant de la difficulté de statuer sur la présence ou non de cet indicateur indirect de fracture. Les hémarthroses isolées étaient fréquentes dans notre étude, représentant 27,2% des radiographies anormales et 12,7% de l'ensemble des examens. Cette fréquence est comparable à la littérature : 28.7% d'hémarthroses pour England (20), 15.7% d'hémarthroses isolées pour Smith (30), 13% pour Rayan (21). La pertinence de considérer une hémarthrose isolée comme une radiographie anormale peut se discuter. En effet, plusieurs études ont montré une faible proportion de fractures lors du suivi des hémarthroses isolées, de l'ordre de 17% (2). Cependant, certaines études retrouvent une proportion plus importante pouvant aller jusqu'à 77% (3). Ces données discordantes de la littérature ainsi que le risque de perte de chance pour une fracture non traitée justifient la prise en charge communément admise, à savoir une immobilisation par plâtre branchio-anté-brachio-palmaire (BABP) avec une consultation et une radiographie de suivi afin d'établir le diagnostic final (1). Il n'existe pas de recommandation sur la durée d'immobilisation et le délai de suivi, témoignant de la difficulté de statuer sur cette question.

Cette étude présentait plusieurs forces, notamment méthodologiques en suivant les recommandations de la CLAIM check-list élaborées en 2020 afin de garantir la qualité et l'homogénéité des études dans le domaine novateur de l'utilisation de l'IA dans l'imagerie médicale (23). Le « ground truth » a été défini avec précision par des experts balayant plusieurs spécialités de la sphère pédiatrique : urgentiste pédiatre, radiologue et orthopédiste infantile, avec un accord fort. L'exhaustivité des inclusions des enfants avec un traumatisme du coude sur une période de plus d'un an limitait les biais de sélection. Si la rigidité des critères d'exclusion des études de validation externe dans la littérature a été critiquée (28,29), leur quasi-absence dans cette étude, avec seulement 1,3% de patients exclus, devrait permettre de généraliser nos résultats à la population générale. De plus, peu d'études publiées analysent des examens composés de deux radiographies face et profil (22), ce qui est une approche plus fidèle à la pratique clinique que l'analyse d'une face ou d'un profil seuls (12,14).

Cette étude comportait plusieurs limites. Elle était rétrospective, et son caractère monocentrique a pu générer des biais de sélection pouvant gêner la généralisation des résultats. Les experts ont analysé les examens en aveugle du contexte clinique, effectuant un diagnostic radiologique seul, ce qui diffère de la pratique en vie réelle des médecins qui ont l'habitude de s'appuyer sur la clinique pour effectuer un diagnostic clinico-radiologique. Les données concernant la prise en charge des patients n'ont pas été recueillies et l'impact potentiel du logiciel sur ces prises en charge n'a pas été étudié.

Conclusion

L'IA est un outil en pleine expansion dans le domaine de l'imagerie médicale avec de nombreux progrès effectués en quelques années, y compris dans la sphère pédiatrique malgré les difficultés spécifiques à cette population. L'IA présente de nombreux avantages, comme la rapidité de ses résultats, sa reproductibilité et son indépendance aux facteurs extérieurs. Hors soin, l'aide d'une IA a montré une amélioration de la lecture des radiographies par des urgentistes et des radiologues (7,22,33). Elle semble permettre l'apport d'une seconde opinion fiable qui perfectionne la détection de fracture par l'humain en attirant son attention sur les radiographies pouvant présenter des anomalies. De plus, des entrainements supplémentaires sur des populations particulières, notamment sur les radiographies de jeunes enfants, permettraient d'améliorer encore ses performances. Néanmoins, l'impact de l'emploi d'une IA sur les prises en charge cliniques a été encore peu étudié (19,29). Une étude prospective en milieu de soin permettrait d'explorer son apport en pratique clinique, tant sur l'impact médico-économique que sur la qualité de vie du patient.

Bibliographie

1. Jouve JL. Guide pratique des urgences en orthopedie pediatrique. 3ème édition. Sauramps Medical; 2015. (Specialites Medicales).

2. Donnelly LF, Klostermeier TT, Klosterman LA. Traumatic elbow effusions in pediatric patients: are occult fractures the rule? Am J Roentgenol. juill 1998;171(1):243-5.

3. Major NM, Crawford ST. Elbow effusions in trauma in adults and children: is there an occult fracture? AJR Am J Roentgenol. févr 2002;178(2):413-8.

4. Wei CJ, Tsai WC, Tiu CM, Wu HT, Chiou HJ, Chang CY. Systematic analysis of missed extremity fractures in emergency radiology. Acta Radiol Stockh Swed 1987. sept 2006;47(7):710-7.

5. Sukvanich P, Samun P, Kongmalai P. Diagnostic accuracy of the shaft-condylar angle for an incomplete supracondylar fracture of elbow in children. Eur J Orthop Surg Traumatol. déc 2019;29(8):1673-7.

6. Hyatt BT, Schmitz MR, Rush JK. Complications of Pediatric Elbow Fractures. Orthop Clin North Am. 1 avr 2016;47(2):377-85.

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. Guly HR. Diagnostic errors in an accident and emergency department. Emerg Med J EMJ. juill 2001;18(4):263-9.

9. Gyftopoulos S, Lin D, Knoll F, Doshi AM, Rodrigues TC, Recht MP. Artificial Intelligence in Musculoskeletal Imaging: Current Status and Future Directions. Am J Roentgenol. sept 2019;213(3):506-13.

10. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. mai 2018;73(5):439-45.

11. Kalmet PHS, Sanduleanu S, Primakov S, Wu G, Jochems A, Refaee T, et al. Deep learning in fracture detection: a narrative review. Acta Orthop. 3 mars 2020;91(2):215-20.

12. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs: Deep learning algorithms—are they on par with humans for diagnosing fractures? Acta Orthop. 2 nov 2017;88(6):581-6.

13. Chung SW, Han SS, Lee JW, Oh KS, Kim NR, Yoon JP, et al. Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. Acta Orthop. 4 juill 2018;89(4):468-73.

14. Gan K, Xu D, Lin Y, Shen Y, Zhang T, Hu K, et al. Artificial intelligence detection of distal radius fractures: a comparison between the convolutional neural network and professional assessments. Acta Orthop. 4 juill 2019;90(4):394-400.

15. Kachalia A, Gandhi TK, Puopolo AL, Yoon C, Thomas EJ, Griffey R, et al. Missed and delayed diagnoses in the emergency department: a study of closed malpractice claims from 4 liability insurers. Ann Emerg Med. févr 2007;49(2):196-205.

16. Hallas P, Ellingsen T. Errors in fracture diagnoses in the emergency department--characteristics of patients and diurnal variation. BMC Emerg Med. 16 févr 2006;6:4.

17. Dupuis M, Delbos L, Veil R, Adamsbaum C. External validation of a commercially available deep learning algorithm for fracture detection in children. Diagn Interv Imaging. mars 2022;103(3):151-9.

18. Hayashi D, Kompel AJ, Ventre J, Ducarouge A, Nguyen T, Regnard NE, et al. Automated detection of acute appendicular skeletal fractures in pediatric patients using deep learning. Skeletal Radiol. 6 mai 2022;

19. Shelmerdine SC, White RD, Liu H, Arthurs OJ, Sebire NJ. Artificial intelligence for radiological paediatric fracture assessment: a systematic review. Insights Imaging. 3 juin 2022;13(1):94.

20. 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. Am J Roentgenol. déc 2018;211(6):1361-8.

21. 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. janv 2019;1(1):e180015.

22. 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.

23. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell. 25 mars 2020;2(2):e200029.

24. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20:37-46.

25. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. mars 1977;33(1):159-74.

26. Lagrange, Rigault. Les fractures de l'extrémité inférieure de l'humérus de l'enfant. 1962;

27. Delacour H, François N, Servonnet A, Gentile A, Roche B. Les rapports de vraisemblance : un outil de choix pour l'interprétation des tests biologiques. Immuno-Anal Biol Spéc. 1 avr 2009;24(2):92-9.

28. Cohen JF, McInnes MDF. Deep Learning Algorithms to Detect Fractures: Systematic Review Shows Promising Results but Many Limitations. Radiology. juill 2022;304(1):63-4.

29. Kuo RYL, Harrison C, Curran TA, Jones B, Freethy A, Cussons D, et al. Artificial Intelligence in Fracture Detection: A Systematic Review and Meta-Analysis. Radiology. juill 2022;304(1):50-62.

30. Smith DN, Lee JR. The radiological diagnosis of posttraumatic effusion of the elbow joint and its clinical significance: the « displaced fat pad » sign. Injury. nov 1978;10(2):115-9.

31. Starosolski ZA, Kan JH, Annapragada A. CNN-based detection of distal tibial fractures in radiographic images in the setting of open growth plates. In: Medical Imaging 2020: Computer-Aided Diagnosis [Internet]. SPIE; 2020 [cité 21 juin 2022]. p. 855-62. Disponible sur: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11314/113143M/CNN-based-detection-of-distal-tibial-fractures-in-radiographic-images/10.1117/12.2549297.full

32. Guermazi A, Tannoury C, Kompel AJ, Murakami AM, Ducarouge A, Gillibert A, et al. Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence. Radiology. mars 2022;302(3):627-36.

33. 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. 6 nov 2018;115(45):11591-6.

Serment médical

Au moment d'être admise 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.

Admise dans l'intimité des personnes, je tairai les secrets qui me seront confiés. Reçue à l'intérieur des maisons, je respecterai les secrets des foyers et ma conduite ne servira pas à corrompre les mœurs.

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 déshonorée et méprisée si j'y manque.

Vu, la Présidente du Jury,

Professeure Christèle Gras-Leguen

Vu, les Directrices de Thèse,

Docteure Bénédicte Vrignaud

Docteure Fleur Lorton

Vu, le Doyen de la Faculté,

NOM : DA COSTA **PRENOM :** Julie

Titre de Thèse : Validation externe des performances diagnostiques d'un logiciel d'intelligence artificielle pour le diagnostic radiologique des fractures du coude de l'enfant

---

RESUME

Introduction : Le *deep learning*, l'une des technologies principales de l'intelligence artificielle, est de plus en plus développé comme outil d'aide au diagnostic pour les cliniciens dans la lecture de radiographie et notamment en traumatologie. Cependant, les études évaluant les performances de ces logiciels pour le diagnostic de fracture dans la population pédiatrique font encore défaut dans la littérature.

Objectif : Réaliser une validation externe des performances du logiciel d'intelligence artificielle Boneview pour le diagnostic des fractures du coude dans une population pédiatrique.

Méthodes : Tous les enfants âgés de 0 à 15 ans et 3 mois ayant consulté aux urgences pédiatriques du CHU de Nantes entre le 1er janvier 2019 et le 1er avril 2020 pour traumatisme du coude et pour lesquels une paire de radiographies a été réalisée ont été inclus rétrospectivement. Le « ground truth », c'est-à-dire le diagnostic radiologique de référence, a été constitué par deux experts (urgentiste pédiatre et radiologue) indépendamment et à l'aveugle du résultat du logiciel et selon deux modalités : normal ou anormal. En cas de désaccord, la radiographie était relue par un troisième expert. Chaque radiographie était classée selon trois modalités par le logiciel : anormale, doute ou normale. Les modalités «anormale » et « doute » ont été regroupées dans nos analyses. Les performances diagnostiques du logiciel (sensibilité, spécificité, valeurs prédictives positive et négative) ont été mesurées.

Résultats : Nous avons inclus 757 enfants (âge médian 8,3 ans) avec une prévalence d'examens radiologiques du coude anormaux de 46,6% ($IC_{95\%}$ 43,1-50,2). Parmi les 757 examens analysés, le logiciel en a classé 54,2% (n=410) en anormaux, 38,2% (n=289) en normaux et 7,7% (n=58) en doute. Le logiciel Boneview avait une sensibilité de 98,0% ($IC_{95\%}$: 96,0 – 99,2), une spécificité de 69,8% ($IC_{95\%}$: 65,1 – 74,2), une valeur prédictive positive de 73,9% ($IC_{95\%}$: 71,0 – 76,7) et négative de 97,6% ($IC_{95\%}$: 95,1 – 98,8).

Conclusion : Le logiciel Boneview a présenté une bonne sensibilité dans une population d'enfants avec traumatisme du coude consultant aux urgences pédiatriques. Des études d'impact devront désormais être menées afin d'évaluer le bénéfice réel pour le patient de l'utilisation en routine de ce logiciel.

MOTS-CLES

Radiographs, Deep learning algorithm, Artificial intelligence, Fractures, Pediatric, Validation studies