UNIVERSITÉ DE LILLE

FACULTÉ DE MÉDECINE HENRI WAREMBOURG

Année : 2022

THÈSE POUR LE DIPLÔME D'ÉTAT

DE DOCTEUR EN MÉDECINE

Évaluation de l'impact d'un algorithme d'intelligence artificielle sur la détection par les radiologues des fractures du poignet et de la main en radiographie

Présentée et soutenue publiquement le 21 Juin 2022 à 18h

Au Pôle Recherche

Par Nicolas CARDOT

JURY

Président :

Madame le Professeur Anne COTTEN

Assesseurs :

Monsieur le Professeur Christophe CHANTELOT

Monsieur le Professeur Xavier DEMONDION

Monsieur le Professeur Eric WIEL

Directeur de thèse :

Monsieur le Docteur Thibaut JACQUES

AVERTISSEMENT

La Faculté n’entend donner aucune approbation aux opinions émises dans les thèses : celles-ci sont propres à leurs auteurs

TABLE DES MATIERES

ABRÉVIATIONS

AUROC : Area Under The Receiver Operating Characteristic (Aire sous la courbe ROC)

CE : Conformité Européenne

DCNN : Deep Convolutional Neural Network (Réseaux de neurones convolutionnels profonds)

DICOM : Digital Imaging and Communications in Medicine (Imagerie numérique et Communication en Médecine)

FDA : Food and Drug Administration

FPN : Feature Pyramid Network

FROC: Free Response Receiver-Operator Curve

IA : Intelligence Artificielle

ML : Machine Learning

PACS: Picture Archiving and Communication System (Système d'archivage et de transmission d'images)

ROC : Receiver-Operator characteristic

ROI : Region of Interest (Zone d'intérêt)

RPN : Region Proposal Network

Se : Sensibilité

Sp : Spécificité

VPN : Valeur Prédictive Négative

VPP : Valeur Prédictive Positive

ÉTAT DES CONNAISSANCES

I - ÉPIDEMIOLOGIE DES FRACTURES DU POIGNET ET DE LA MAIN

Le nombre de passages annuels aux urgences en France est en constante augmentation, en effet ce dernier s'établissait à 10 millions en 1996 et a augmenté progressivement d'environ 3.3% par an pour atteindre 21 millions en 2019 (1,2).

Les fractures sont un des principaux motifs de consultation aux urgences dans les pays occidentaux, peuvent être responsables d'une diminution importante de la qualité de vie perçue chez les patients (3) et engendrent un retentissement économique majeur : les traumatismes de la main et du poignet ont par exemple un coût annuel estimé à 700 millions d'euros par an aux Pays Bas (4).

Les fractures du radius distal sont les fractures les plus fréquentes du squelette appendiculaire et représentent une fracture sur six prise en charge aux urgences, avec deux pics de fréquence : le sujet âgé et ostéoporotique après un traumatisme mineur et le sujet jeune après un traumatisme violent (5) (6).

Les fractures du carpe représentent 3% des fractures de l'adulte et peuvent facilement passer inaperçues. En l'absence de traitement, elles peuvent être à l'origine d'une instabilité chronique, de douleurs résiduelles ou de raideur. La fracture du scaphoïde représente la fracture la plus fréquente des os du carpe (70 à 90%), suivie des fractures du triquetrum (12%) (7).

Devant une suspicion clinique de fracture, le diagnostic repose en première intention sur les radiographies à l'aide d'au minimum deux incidences orthogonales (face et profil) auxquelles on peut adjoindre des clichés supplémentaires (trois quart, cliché

comparatif) en raison de l'accessibilité de l'examen, de sa facilité de réalisation et de sa faible irradiation.

Cependant il a été démontré que les radiographies seules ont malgré tout des performances imparfaites pour la détection des lésions post-traumatiques, notamment de la main et du poignet. La sensibilité des radiographies pour la détection des fractures du radius et de l'ulna est estimée entre 72.8% et 80% (8). *Balci et al.* ont montré que la sensibilité des radiographies standard pour la détection des fractures du carpe est de 38.7% (8). La sensibilité des radiographies standard pour la détection des fractures du scaphoïde, qui sont les fractures les plus communes du carpe, est estimée entre 67% et 80% (5)(8)(9).

Les fractures non détectées représentent jusqu'à 79% des erreurs médicales dans les services d'urgences, et parmi ces fractures les plus représentées sont les fractures du poignet de de la main (10).

Dans la plupart des cas, les radiographies réalisées aux urgences sont examinées dans un premier temps par des médecins non spécialisés dans l'interprétation d'examens de radiologie. Ceci est d'autant plus vrai pour les examens réalisés au cours de la permanence des soins, la nuit et les week-ends. Il a en effet été démontré que le taux d'erreurs diagnostiques est plus élevé lors de ces périodes, du fait de la difficulté à obtenir un avis spécialisé et de la fatigue accumulée (11).

Face à ces défis actuels et croissants, l'intelligence artificielle (IA) et notamment l'apprentissage automatique (*Machine Learning* - ML) pourraient représenter une piste d'optimisation de la prise en charge des patients, en fournissant une assistance aux

médecins afin de maintenir ou même d'améliorer la précision du diagnostic, malgré l'augmentation du volume d'examens à interpréter. Les radiologues sont d'ailleurs en majorité demandeurs de ce type d'outils, afin de limiter le risque d'erreurs médicales et d'augmenter la fiabilité de leurs interprétations (12).

II - INTELLIGENCE ARTIFICIELLE

1 - Définitions

L'intelligence artificielle correspond à l'ensemble des théories et des techniques développant des programmes informatiques complexes capables de simuler certains traits de l'intelligence humaine (13).

L'idée de l'intelligence artificielle est née dans les années 1950, notamment du fait des travaux d'Alan Turing, et de la création de son « Imitation Game » aussi appelé « Test de Turing » permettant de déterminer si une machine est « intelligente ».

Depuis, l'intelligence artificielle s'est considérablement développée. Le terme regroupe plusieurs concepts et constitue un champ très large de recherche.

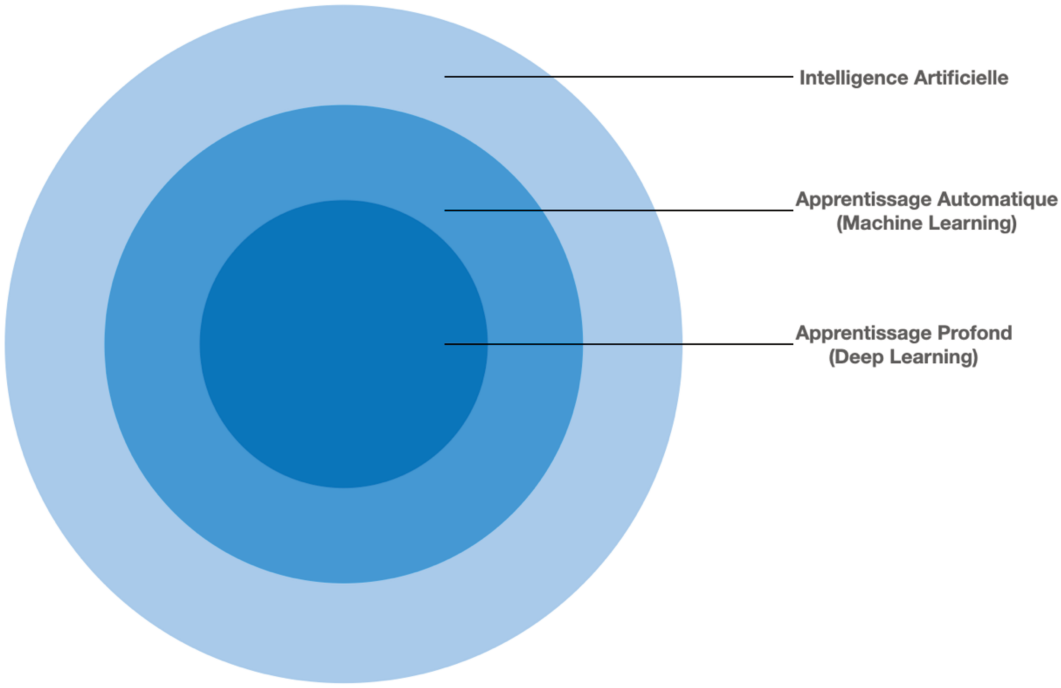


Figure 1 : Sous types d'intelligence artificielle

Le *Machine Learning* (ML) ou apprentissage automatique correspond à un champ d'étude de l'intelligence artificielle qui permet aux ordinateurs « d'apprendre » à partir de données, et d'améliorer leurs performances à résoudre des tâches. Les algorithmes d'apprentissage automatique construisent un modèle en se basant sur des exemples de données, appelées données d'apprentissage, afin de produire des prédictions ou des décisions sans être explicitement programmés pour cette tâche.

Il existe trois principaux sous-types d'apprentissage automatique :

- L'apprentissage supervisé

Les algorithmes d'apprentissage supervisé construisent un modèle mathématique à partir d'un ensemble de données labélisées (données d'entrainement) afin de pouvoir

ensuite répartir les nouvelles données d'entrée parmi un ensemble de classes définies en sortie. Grâce à l'optimisation itérative du modèle mathématique, les algorithmes d'apprentissage supervisé apprennent une fonction qui peut donc être utilisée pour classer des nouvelles données selon un ensemble de classes prédéfinies.
Il s'agit du type d'apprentissage automatique le plus fréquemment utilisé actuellement.

- L'apprentissage non-supervisé

Les algorithmes d'apprentissage non supervisé apprennent à partir de données qui n'ont pas été labélisées et au lieu de répondre aux « *feedback* » du modèle mathématique, identifient les points communs dans les données afin de les regrouper en « *clusters* ». Cette méthode d'apprentissage automatique nécessite une quantité de données nettement supérieure à l'apprentissage supervisé, reste actuellement d'application limitée sur les données de santé et est surtout utilisée en statistiques, afin d'extraire des points communs et des liens au sein de jeux de données (14).

- L'apprentissage par renforcement

L'apprentissage par renforcement est un domaine de l'apprentissage automatique qui permet à un algorithme d'apprendre à partir d'interactions avec le milieu dans lequel il est plongé. L'algorithme prend des décisions en fonction de son état courant, et à chaque interaction avec son milieu, il reçoit un signal rétroactif positif ou négatif afin d'adapter ses décisions.

Le *Deep Learning* ou Apprentissage Profond est une forme de *Machine Learning* par apprentissage automatisé basé sur l'utilisation de réseaux de neurones artificiels profonds (*deep convolutional neural networks – DCNN*).

Il s'agit de réseaux constitués de plusieurs couches de neurones interagissant entre eux de manière antérograde et rétrograde en ayant le but de s'approcher du mode de fonctionnement du cerveau humain. Le *Deep Learning* existe depuis les années 1980 mais a connu un véritable essor dans les années 2010 grâce au progrès de la technologie, et notamment de l'augmentation des capacités de traitement du signal informatique. Il s'agit à l'heure actuelle de la méthode d'apprentissage automatique la plus utilisée pour la création d'algorithmes d'intelligence artificielle en radiologie.

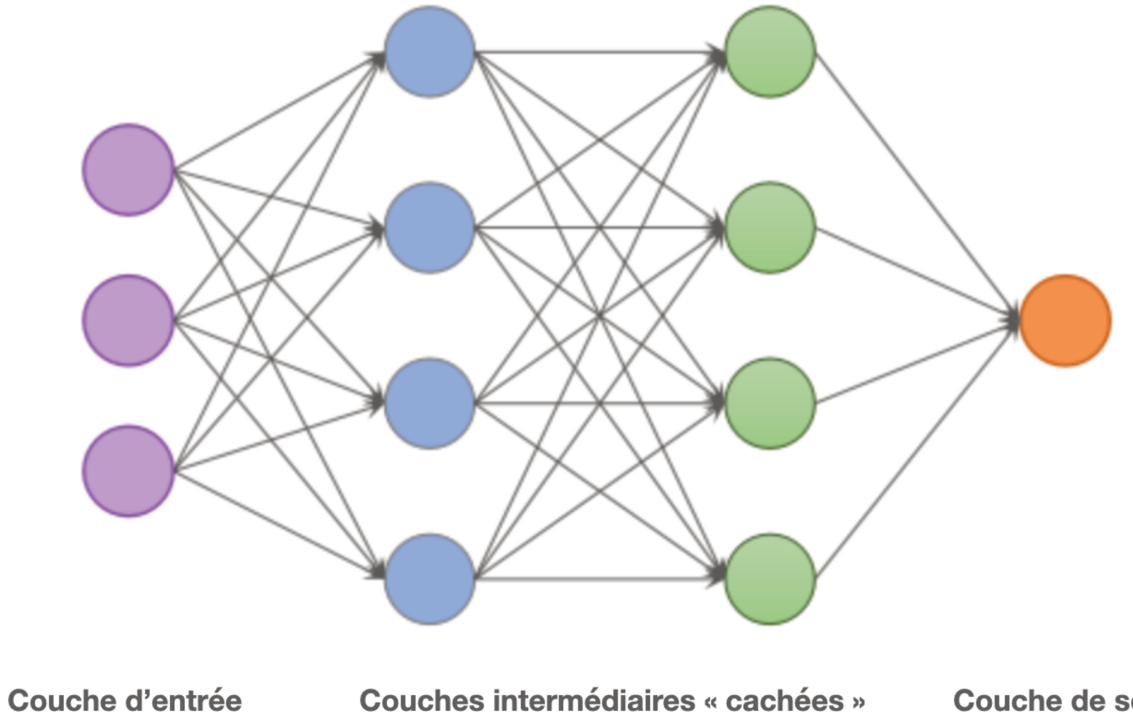


Figure 2 : Représentation d'un réseau de neurones artificiels

2 - Développement d'un modèle

Afin de créer un algorithme d'intelligence artificielle, il est nécessaire de disposer d'un grand nombre de données qui constituent l'élément central du développement. Dans le cadre d'un apprentissage supervisé, ces données doivent avoir été « labelisées »

afin d'établir le « Ground Truth » qui constituera la vérité considérée absolue à partir de laquelle sera développé le modèle.

Ces données vont être divisées en trois ensembles :

- Les données d'entrainement qui vont servir à entrainer le modèle. Elles doivent être représentatives de la cible, et les différentes classes doivent être idéalement équilibrées. Pour les tâches de classification, un algorithme d'apprentissage supervisé va examiner l'ensemble des données d'entrainement pour déterminer les combinaisons optimales de variables qui généreront un bon modèle prédictif (15).

- Les données de validation qui vont servir à réaliser une première évaluation du modèle afin d'affiner sa précision après sa phase d'entrainement, par de nouvelles itérations permettant d'ajuster des hyperparamètres, afin de sélectionner le modèle présentant les meilleures performances (15).

- Le jeu de donnée de test qui constitue un ensemble de données indépendantes de l'ensemble des données d'entrainement et de validation et qui va servir uniquement à évaluer les performances finales du modèle qui a été sélectionné à la fin de la phase de validation (15).

Cette phase de test interne (basée sur des données proches de celle ayant servi à l'entraînement) doit idéalement être complétée par une phase de test externe (sur des données provenant d'une autre cohorte, d'un autre centre) afin d'avoir une meilleure représentation des performances réelles du modèle. Cette phase est malheureusement absente de plus de 90% des études (16).

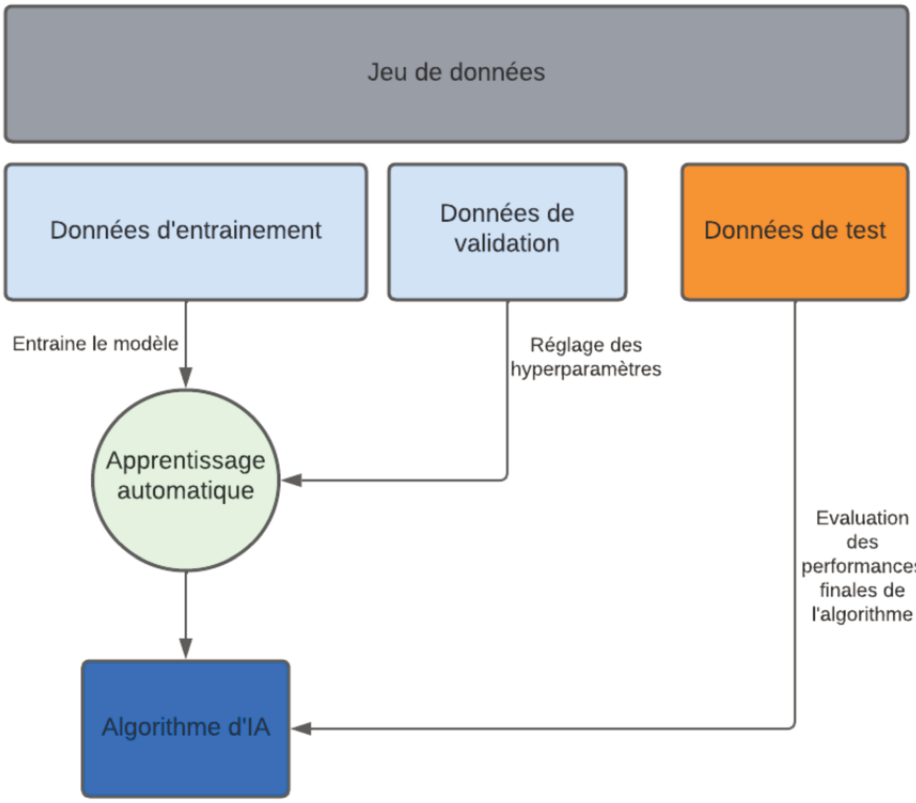


Figure 3 : Représentation des données, divisées en trois sous-ensembles

3 - Généralisabilité d'un modèle d'IA

Le défi central de l'apprentissage automatique est que le modèle développé doit fonctionner de manière équivalente entre des données inédites et les données sur lesquelles il a été développé. Cette capacité est appelée la généralisabilité.

Elle est déterminée par deux caractéristiques,

- Le biais, aussi appelé erreur d'entrainement, qui correspond au taux d'erreur de prédiction obtenu lorsque le modèle est appliqué aux données sur lesquelles il a été entrainé.

- La variance, aussi appelé erreur de généralisation, qui correspond à l'écart de performance du modèle sur les données d'entrainement et sur les données de test.

Ces deux paramètres correspondent à deux enjeux majeurs de l'apprentissage automatique :

- Le sous-ajustement (*under-fitting*) qui se produit lorsque le modèle n'est pas en mesure d'obtenir une erreur d'entrainement suffisamment faible. Le modèle n'est pas assez entrainé et est alors trop simple et n'est pas généralisable.
- Le surajustement (*over-fitting*) qui se produit lorsque le modèle correspond trop étroitement aux données sur lesquelles il a été entrainé. Il en résulte un modèle avec d'excellent résultats sur les données d'entrainement (biais faible) mais qui n'est pas généralisable à des données extérieures (variance élevée).

Il s'agit donc d'un compromis à faire entre le biais et la variance afin d'obtenir le modèle le plus généralisable possible (dilemme biais-variance) (15)

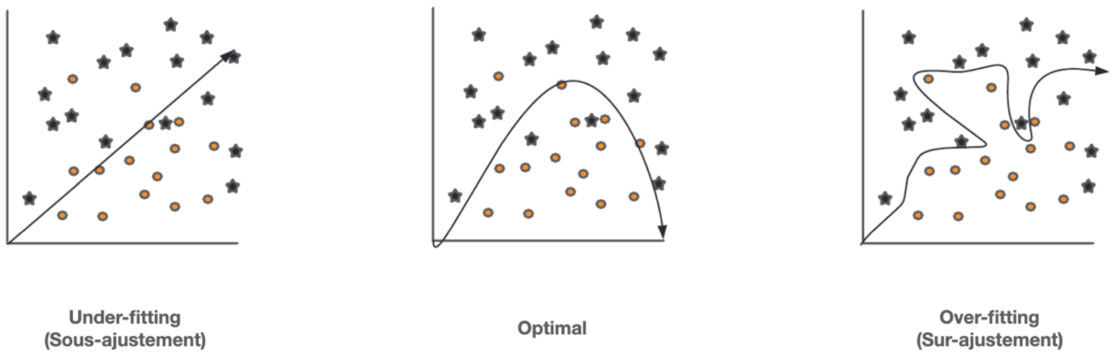


Figure 4 : Représentation du sous-apprentissage et du sur-apprentissage

III - APPLICATIONS AU DOMAINE DE LA SANTÉ

1 - D'un point de vue général

Les premières applications de l'IA en médecine se sont concentrées en grande partie sur les tâches de diagnostic et de reconnaissance d'image telles que détecter la rétinopathie chez les personnes diabétiques grâce à des photographies de fond d'œil, détecter des lésions en mammographie ou reconnaitre les lésions dermatologiques suspectes de malignité (17)(18)(19)

Des applications de l'intelligence artificielle en imagerie médicale sont possibles, telles que le tri des indications, l'amélioration et l'optimisation de l'acquisition des images ainsi que la réduction de la dose (20), l'optimisation du flux des patients (21) ou la création de compte rendus automatisés. Elles sont également possibles pour proposer une aide au diagnostic, ce qui constitue le principal axe de recherche actuellement.

2 - Application en imagerie musculo-squelettique

L'avènement de l'IA, et plus particulièrement de l'apprentissage automatique, a donné lieu à un nombre croissant de publications sur les applications potentielles dans de nombreux contextes, notamment en imagerie musculo-squelettique.

L'automatisation du diagnostic des fractures a été l'objectif principal des recherches en IA sur l'imagerie traumatique musculo-squelettique.

Les études menées jusqu'à présent se sont principalement concentrées sur la détection de lésions post-traumatiques, et notamment sur des lésions particulièrement courantes, telles que les fractures vertébrales (22)(23)(24)(25)(26), les fractures du bassin et de la hanche (27)(28)(29)(30)(31), les fractures costales (32)(33)(32)(34), ou les fractures du poignet et de la main (35)(36)(37)(38)(39)(40).

La *Food And Drug Administration* (F.D.A.) a autorisé à ce jour (*mai 2022*) l'utilisation de 188 solutions d'intelligence artificielle en santé, dont 13 pour l'imagerie musculo-squelettique et 4 pour la recherche de fractures (41).

Ces algorithmes étant marqués comme des aides à la décision médicale (et non pas des dispositifs autonomes), leur responsabilité propre et les publications requises sont principalement basées sur la sécurité des données. Les données d'efficacité ne nécessitent pas forcement d'avoir été publiées dans des revues à comité de lecture. Les études d'efficacité (de qualité variable) suivent cependant deux modèles habituels.

A) Modèle d'étude « IA contre Radiologue »

Les premières études sur cette thématique ont été souvent construites de façon à évaluer les performances brutes d'un algorithme ou de l'évaluer en comparaison avec des lecteurs humains. La plupart des algorithmes publiés dans la littérature démontrent une grande précision dans l'identification des fractures, par exemple Kim *et al.* ont rapporté dans leur étude une sensibilité de 90% pour la détection des fractures du radius distal (42). Thian *et al.* rapportent quant à eux une sensibilité d'environ 90% pour la détection des lésions du radius et de l'ulna (39).

Lors de la comparaison avec les professionnels de santé, les algorithmes rapportent en général des performances non significativement inférieures voir supérieures à celles des experts, tels que les radiologues ayant bénéficié d'une spécialisation en imagerie musculosquelettique ou les chirurgiens orthopédiques. En revanche ces algorithmes rapportent souvent de meilleures performances diagnostiques que les

lecteurs non experts, qui sont pourtant fréquemment les premiers à visualiser et interpréter ces examens (43)(44).

Cheng *et al.* rapportent un modèle de détection et de classification des fractures de la hanche qui a des performances diagnostiques non inférieures à celles de lecteurs experts tels que les radiologues spécialisés en imagerie musculo squelettique, et des performances supérieures à celles des médecins urgentistes (28).

Gan *et al.* ont développé un modèle d'IA de détection des fractures du radius distal qui rapporte des performances de détection équivalentes à celles de chirurgiens orthopédistes et supérieures à celles de radiologues non spécialisés en imagerie musculosquelettique (40).

Ces études ont cependant deux limites principales :

- Les performances de l'algorithme sont généralement sur-évaluées, car elles sont calculées sur des données de test interne, habituellement très proches des données ayant servi à l'entrainement (16), entraînant *de facto* une surévaluation du modèle dans environ 81% des cas (45).
- Les performances des lecteurs humains sont généralement sous-estimées, il a en effet été démontré que les performances de détection humaine sont inférieures sur un jeu de données rétrospectif, d'autant plus si ce dernier est enrichi (avec une prévalence supérieure à la pratique courante), par rapport aux performances observées sur un jeu de données similaire en pratique clinique prospective (« *laboratory effect* ») (46)

B) Modèle d'étude « Radiologue augmenté par l'IA »

Dans un second temps les études ont été plutôt développées dans le but d'évaluer « le radiologue augmenté par l'intelligence artificielle » en comparant les performances d'un radiologue seul par rapport à un radiologue aidé de l'intelligence artificielle. Il a été démontré à plusieurs reprises que les algorithmes d'IA pourraient permettre d'améliorer les performances de détection des fractures par les médecins.

Dans leur article, Guermazi *et. al*. ont étudié l'impact de l'IA sur les performances diagnostiques de 24 médecins d'horizons et de spécialités variés, interprétant des radiographies. Les auteurs ont constaté que l'IA a amélioré la sensibilité des médecins dans le diagnostic des fractures (passant de 64.8% à 75.2%) et que la spécificité des lecteurs s'est améliorée en passant de 90.6 % à 95.6 % dans le même temps (43). Duron *et. al.* ont utilisé une méthodologie similaire avec un groupe de 12 médecins (6 radiologues et 6 urgentistes) et rapportent une augmentation de la sensibilité de l'ordre de +8.7% grâce à l'aide de l'IA (44).

Lindsey *et. al.* ont démontré que les médecins urgentistes pourraient atteindre un niveau de sensibilité de 95% pour la détection des fractures du poignet et de la main lorsqu'ils sont assistés par des algorithmes d'IA, ce qui permettrait de réduire le taux d'interprétations erronées de 47% (38). Sato *et. al.* ont rapporté un modèle de détection qui a permis d'améliorer significativement la détection des fractures de hanche par des internes en médecine n'ayant pas encore choisi leur spécialité (30).

Ces résultats soulignent l'idée que les algorithmes d'apprentissage automatique pourraient permettre d'améliorer la précision des professionnels de santé pour la détection des fractures.

3 - Limites des études actuelles

Mais alors que les algorithmes ML pour l'évaluation des traumatismes musculo-squelettiques ont fait de grands progrès, des limites demeurent.

Peu d'études ont évalué les DCNN sur l'ensemble des fractures visibles sur les radiographies et se sont plutôt concentrées sur une fracture définie et en règle générale courante, telles que les fractures de la hanche (28), ou les fractures du radius (38). Il n'est donc pas possible de généraliser ces modèles à la détection de l'ensemble des fractures du corps humain sans études complémentaires. Quelques études ont néanmoins évalué des algorithmes d'IA sur un panel varié de fractures et ont rapportés de bonnes performances diagnostiques (43)(44).

Peu d'algorithmes de détection des fractures en radiographie ont été soumis à un test de validation externe dont le but est d'évaluer la généralisabilité du modèle en évaluant des biais potentiels tels que le surajustement ou la stratification cachée. Il s'agit d'une étape cruciale avant l'implantation en pratique clinique, les modèles d'IA ayant, en général, tendance à avoir de moins bonnes performances sur les test de validation externe que sur les données utilisées pour créer le modèle (47).

Yu *et. al.* ont réalisé une revue de la performance de modèles d'IA lors de test de validation externe et ont estimé que la majorité des 86 modèles testés (81%) présentaient une baisse de performances par rapport au test de validation interne (45). Or, Carmo *et. al.* ont démontré dans une méta analyse des études rapportant la création d'un algorithme d'IA pour la recherche des fractures, que seulement 4 des 36 études analysées rapportaient une forme de validation externe (temporelle ou

géographique), et aucun des algorithmes actuels n'avait été testé de manière prospective (48).

Par ailleurs, les études évaluant les algorithmes pour la détection des fractures le font souvent sur la base d'un « *Ground Truth* » établi par l'annotation des radiographies par un ou plusieurs lecteurs experts (43)(44), mais il a été démontré que les performances diagnostiques de la radiographie standard sont imparfaites (8), avec par exemple une sensibilité diagnostique des radiographies pour les fractures du poignet et de la main estimée à 63% (8).

La labellisation des données est une étape clé dans la création d'un modèle d'IA robuste. Ceci signifie que les données contiennent des sous-ensembles de cas non reconnus qui peuvent engendrer une stratification cachée (« *Hidden Stratification* »).

Il s'agit de la non prise en compte de certaines fractures par les experts lors de la labellisation car non visibles en radiographies. Ceci peuvent affecter la formation du modèle, les performances du modèle et, surtout, les résultats cliniques (49).

Une façon d'évaluer cette « stratification cachée » est de tester les algorithmes à l'aide de jeux de donnée dont le « *ground truth* » a été établi à l'aide d'un critère robuste tel que le scanner ou l'arthroscopie pour les fractures.

Quelques études ont étés menées avec cette méthodologie et rapportent une diminution des performances diagnostiques des modèles.

Raisuddin *et. al.* ont démontré qu'un modèle d'IA entrainé à détecter les fractures du poignet présentait une baisse significative de ses performances diagnostiques sur un jeu de données de validation externe qualifié de difficile, qui avait nécessité un scanner

complémentaire afin d'affirmer le diagnostic, avec des AUROCS passant de 0.99 sur le test classique à 0.84 sur le test difficile (50).

Gipson *et. al.* ont évalué les performances d'une solution d'intelligence artificielle (commercialisée) pour l'interprétation des radiographies de thorax post-traumatiques en utilisant comme « *Ground Truth* » le scanner qui avait été réalisé dans les suites immédiates. Elle rapporte d'excellentes performances de l'IA, équivalentes à celles des radiologues voir supérieures en ce qui concerne la détection de pneumothorax et d'atélectasie. Une des implications est que si le « *Ground truth* » pour ces anomalies avait été annoté par des experts sur des radiographies, la sensibilité réelle de détection des lésions post-traumatiques thoraciques par l'algorithme aurait été mal évaluée (33).

Ozkaya *et. al.* ont évalué les performances d'un algorithme d'IA pour la détection des fractures du scaphoïde en radiographie, en utilisant comme gold standard les données d'un scanner. Ils rapportent une sensibilité de 76% et une spécificité de 92%. Ils ont réalisé une comparaison de cet algorithme avec les performances de deux chirurgiens orthopédiques et d'un médecin urgentiste, retrouvant des performances de l'algorithme équivalentes à celles des orthopédistes, supérieures à celles de l'urgentiste (51). Cette étude portait cependant sur un algorithme développé pour l'étude (non validé et non commercialisé) et sur une faible cohorte de lecteurs (n=3).

Peu d'études ont été réalisées en condition prospective, en effet Carmo *et. al.* rapportent dans une méta analyse que sur les 36 algorithmes de détection des fractures qui ont été évalués, aucun ne l'a été de manière prospective (48).

En effet la plupart des études sont rétrospectives et les médecins testés ne sont pas forcément dans des conditions de travail optimales (*laboratory effect*) (46) et sont habituellement en aveugle des renseignements cliniques, ce qui peut potentiellement diminuer leurs performances (52).

Par exemple, il a été démontré que les informations cliniques comme la topographie du traumatisme ou son mécanisme améliorent les performances des radiologues dans la détection d'une fracture, en augmentant notamment leur confiance diagnostique (53). Par exemple, l'utilisation de schémas pour localiser la zone du traumatisme fait passer la sensibilité de détection de fractures subtiles de 67% à 73%, et la spécificité de 93% à 94% (54).

Enfin, les correspondants médicaux des radiologues ne se sentent pas nécessairement à l'aise pour prendre des décisions médicales basées uniquement sur des rapports de radiologie émis de façon autonome par un modèle d'IA. En revanche, ces correspondants médicaux rapportent être aussi à l'aise avec des compte rendus émis uniquement par des radiologues qu'avec des comptes rendus émis par un radiologue assisté d'un modèle d'IA (55). Les patients quant à eux semblent pour l'instant méfiants vis-à-vis des modèles d'IA et préfèrent l'avis d'un professionnel de santé, assisté ou non d'un modèle d'IA, à celui d'un modèle d'IA seul (56)(57).

Ces éléments montrent que, malgré des premiers résultats prometteurs rapportés dans de nombreux articles, il est nécessaire de valider consciencieusement les modèles d'IA, notamment en utilisant des jeux de données de validation externe, une labellisation à l'aide d'une référence robuste et des conditions satisfaisantes d'interprétation par les lecteurs humains, idéalement de façon prospective. Cela, afin

de permettre à la fois d'obtenir une évaluation précise des performances des modèles d'IA ainsi que de leurs limites, et d'en augmenter la compréhension et l'acceptabilité.

IV - OBJECTIFS DE L'ÉTUDE

L'objectif de l'étude était d'évaluer l'impact d'un algorithme d'intelligence artificielle commercialisé, sur les performances de détection des fractures du poignet et de la main par des radiologues, en se basant sur un « *Ground truth »* non pas radiographique mais scanographique.

L'algorithme utilisé était BoneView™, développé par la société française Gleamer, marqué CE et FDA pour l'aide à la détection de fractures en radiographie standard. Les radiologues ont été évalués de façon rétrospective, sans puis après activation de l'IA.

Il s'agit à notre connaissance de la première étude visant à évaluer l'impact d'un algorithme commercial sur la détection de l'ensemble des fractures présentes sur des radiographies du poignet et de la main en se basant sur un jeu de données entièrement validé par un gold standard scanner.

ARTICLE SCIENTIFIQUE

I - ABSTRACT

INTRODUCTION

Les fractures non détectées sont la cause principale d'erreurs médicales aux urgences, en particulier les fractures du poignet de de la main. Des algorithmes d'intelligence artificielle (I.A.) spécialisés dans l'aide à la détection de fractures en radiographie standard ont été développés ces dernières années et se déploient en pratique clinique. Néanmoins, la référence de ces algorithmes est uniquement radiographique, et peut donc sous-estimer les fractures par rapport à un *ground truth* scanographique, notamment pour des zones anatomiques difficiles. L'objectif de cette étude était d'évaluer les performances de radiologues sans puis après l'aide d'un algorithme d'I.A. (BoneView®) dans la détection des fractures du poignet et de la main en radiographie standard, en utilisant le scanner comme *ground truth*.

MATÉRIEL ET MÉTHODES

Un jeu de données composé de radiographies de la main et du poignet ainsi que des scanners concomitants, réalisés dans un contexte post-traumatique aux urgences d'un centre hospitalier et universitaire, a été constitué de façon rétrospective. Toutes les radiographies ont été annotées par deux radiologues, en consensus, en fonction des anomalies post-traumatiques visibles en scanner. Les données étaient composées de 296 patients, 118 ne présentant aucune fracture (39.9%) et 178 avec au moins une fracture (60.1%) pour un total de 267 fractures visibles en scanner. Vingt-trois radiologues avec des niveaux d'expertise différente (14 radiologues seniors, 9 internes de radiologie) ont été inclus pour analyser rétrospectivement l'ensemble des radiographies, sans puis avec l'aide de l'I.A., en aveugle des résultats du scanner.

RÉSULTATS

Sur les données radiographiques, en se basant sur le *ground truth* scanner, les performances de l'IA pour la détection de fracture étaient les suivantes : 72.5% (sensibilité), 89.8% (spécificité), 68.4% (valeur prédictive négative - VPN), 91.5% (valeur prédictive positive - VPP) et 0.764 (AUROC), ce qui est significativement inférieur aux résultats publiés sur des jeux de données utilisant un *ground truth* radiographique, témoignant d'une potentielle sous-estimation du nombre de fractures manquées dans la majorité des publications n'utilisant pas un référentiel plus robuste que les radiographies. Néanmoins, l'utilisation de l'algorithme a permis aux radiologues d'améliorer leur sensibilité, passant de 65.8% à 70.3% à l'échelle des fractures ($p<0.0001$) et de 58.2% à 63.5% à l'échelle des patients ($p<0.0001$), ainsi que leur VPN, passant de 58.5% à 61.8% ($p<0.0001$), indépendamment de leur niveau d'expertise. La spécificité des lecteurs n'était pas significativement affectée, passant de 88.5% à 89.1% ($p=0.91$) de même que leur VPP, de 88.7% à 89.9% ($p=0.08$).

CONCLUSION

L'utilisation d'un algorithme d'I.A. a permis aux radiologues d'améliorer significativement leur sensibilité et leur valeur prédictive négative pour la détection des fractures du poignet et de la main en radiographie, sans affecter ni leur spécificité ni leur valeur prédictive positive. Ces données confortent des publications récentes sur cette thématique. L'utilisation d'un *ground truth* scanner comme référence est novateur pour ce type de méthodologie, et est à l'origine de performances significativement inférieures de l'algorithme par rapport à celles rapportées dans la littérature, mais probablement plus proches de la réalité clinique.

II - INTRODUCTION

Les fractures non détectées représentent environ 80 % des erreurs médicales dans les services d'urgences et, parmi ces fractures, les plus représentées sont les fractures du poignet de de la main (10). Ce risque d'erreur est d'autant plus important que la charge de travail dans les services d'urgences a fortement augmenté ces dernières années (2). De plus, le taux d'erreurs diagnostiques est plus élevé lors des périodes où une interprétation spécialisée par un radiologue est difficile à obtenir, comme lors des gardes de nuit (11).

Le développement récent d'algorithmes d'intelligence artificielle (IA) appliqués à l'imagerie traumatologique pourrait possiblement permettre d'améliorer les performances de détection des fractures par les médecins. Une étude récente a montré que l'IA pourrait améliorer la sensibilité des médecins dans le diagnostic des fractures (passant de 64.8% à 75.2%) ainsi que leur spécificité (passant de 90.6 % à 95.6%) (43). Une deuxième étude a rapporté une augmentation de la sensibilité des médecins de l'ordre de +8.7% grâce à l'aide de l'IA (44).

Devant un traumatisme du poignet ou de la main, les radiographies standard sont le premier examen réalisé en routine clinique. La sensibilité des radiographies pour la détection des fractures du radius et de l'ulna est estimée au maximum à 80%, et 39% pour la détection de l'ensemble des fractures du carpe (8). Pour la détection des fractures du scaphoïde, qui sont les fractures les plus fréquentes du carpe, cette sensibilité est estimée entre 67% et 80% (5)(8)(9). Or, les études évaluant la performance des algorithmes d'I.A. pour la détection des fractures le font sur la base d'un « *Ground Truth* » établi par l'annotation des radiographies par un ou plusieurs lecteurs experts (43)(44), alors même que cet examen a des performances

imparfaites. Cela pourrait donc entrainer une mauvaise estimation des performances réelles des lecteurs humains, de l'algorithme, et des médecins aidés par l'algorithme. A l'inverse, d'autres examens d'imagerie comme le scanner ont des performances supérieures aux radiographies standard pour le diagnostic de fracture, avec des sensibilités rapportées pouvant aller jusqu'à 100% dans certaines études récentes (58,59).

L'objectif de cette étude était d'évaluer l'impact d'un algorithme d'intelligence artificielle commercialisé, sur les performances de détection des fractures du poignet et de la main par des radiologues, en se basant sur un *ground truth* non pas radiographique mais scanographique.

III - MATÉRIEL ET MÉTHODES

JEU DE DONNÉES DE L'ÉTUDE

Pour réaliser cette étude, un jeu de données d'images a été constitué. Ce dernier contenait les radiographies de patients adultes pris en charge pour un traumatisme de la main ou du poignet dans un centre hospitalier universitaire (CHU de Lille, France), ainsi que les scanners correspondants.

A l'aide d'une requête dans le PACS (Intellispace, Philips), tous les patients ayant bénéficié lors de leur passage aux urgences à la fois d'une radiographie et d'un scanner de la main ou du poignet, entre Janvier 2016 et Août 2019, ont été recensés rétrospectivement. Les cas ont ensuite été dé-identifiés individuellement, après avoir regroupé pour chaque patient les radiographies et le scanner correspondant, permettant ainsi de conserver une table de correspondance après dé-identification.

DÉFINITION DU GROUND TRUTH

Les radiographies dé-identifiées ont été importées dans la plateforme d'annotation de Gleamer (Paris, France). Les scanners dé-identifiés ont été importés dans une visionneuse adaptée (SyngoVia, Siemens Healthineers). Pour chaque patient, à l'aide d'une table de correspondance, les radiographies standard ont ensuite été annotées, à partir du scanner, par deux radiologues, en consensus (8 ans et 3 ans d'expérience en imagerie traumatologique).

Les anomalies retenues comme pathologiques en scanner étaient les fractures récentes et les avulsions osseuses d'aspect récentes. Chaque anomalie post-traumatique visualisée en scanner était annotée sur chaque incidence radiographique, en positionnant une boîte sur la zone pathologique, y compris si la lésion présente en scanner n'était pas visible radiographiquement. Le jeu de données final de

radiographies annotées était donc un reflet de la réalité des fractures et avulsions visibles en scanner.

INTERPRÉTATION PAR LES LECTEURS

Toutes les radiographies du jeu de données ont été interprétées par 23 lecteurs indépendants, dont 9 internes en radiologie ayant déjà une expérience en imagerie traumatologique et 14 radiologues séniors spécialisés en imagerie musculosquelettique (nombre d'années d'expérience en radiologie : 2-25, moyenne : 5.6 ans +/- 4.6), en aveugle des scanners, sans contexte clinique hormis la notion de traumatisme.

L'interprétation a été réalisée sur la plateforme d'annotation de Gleamer (Paris, France), en utilisant des écrans de qualité diagnostique. Les lecteurs disposaient de 3 mois pour réaliser l'ensemble des interprétations, du 1er septembre au 30 novembre 2021.

Pour chaque cas, les lecteurs devaient d'abord analyser toutes les incidences radiographiques sans l'aide de l'intelligence artificielle, en positionnant des labels sur la (les) zone(s) considérée(s) comme pathologique(s), ou aucune boîte en cas de radiographies considérées comme normales, avant de pouvoir activer l'intelligence artificielle. Les annotations des lecteurs étaient enregistrées automatiquement lors de l'activation de l'intelligence artificielle.

A l'aide des résultats proposés par l'algorithme, les lecteurs pouvaient ensuite ajouter, retirer ou modifier les labels en fonction de leur analyse finale. Les annotations finales étaient enregistrées automatiquement à la clôture de chaque cas. Les radiographies étaient présentées dans un ordre aléatoire, différent pour chacun des lecteurs.

Pour chaque lecteur, la sensibilité, spécificité, valeur prédictive positive (VPP) et valeur prédictive négative (VPN), sans puis avec utilisation de l'IA, ont été calculées, en se basant sur le *ground truth* issu du scanner. La sensibilité a été calculée à l'échelle des patients et à l'échelle des fractures individuelles pour mieux évaluer les modifications de détection de chaque fracture, y compris pour les cas présentant plusieurs lésions. Une analyse des performances par sous-groupe a également été réalisée en fonction de l'ancienneté des lecteurs (internes versus radiologues seniors) et en fonction de la topographie de la fracture.

ALGORITHME UTILISÉ

BoneView™ (Gleamer, Paris, France) est un algorithme d'intelligence artificielle d'aide au diagnostic, disponible commercialement. Cet algorithme est un dispositif médical marqué C.E. et F.D.A., dont l'objectif est d'aider à la détection de fractures récentes sur des images DICOM de radiographies en pleine résolution.

L'algorithme est un réseau de neurones convolutionnel profond (DCNN) basé sur le *framework* de détection d'objets « Detectron 2 », qui fonctionne selon un mode de détection en deux étapes. La première étape permet la réception des images DICOM en entrée, sans pré-traitement ni recadrage, puis l'extraction de cartes paramétriques intermédiaires en utilisant un *Feature Pyramid Netword* (FPN). Un *Region Proposal Network* (RPN) traite ensuite ces cartes paramétriques en générant des cibles (*box*), et en leur attribuant chacune un score basé sur la probabilité qu'elle contienne un objet (ici, une fracture). Si le score généré est supérieur à un seuil fixé, cette cible est considérée comme zone d'intérêt (ROI). Les caractéristiques contenues dans cette cible ainsi que ses coordonnées spatiales sont transmises pour la deuxième étape, visant à affiner les résultats du RPN. Quand les scores de confiance dépassent les

seuils de fonctionnement préétablis, l'algorithme fait apparaître la région d'intérêt (ROI) sous la forme d'un rectangle sur les images natives.

L'algorithme a été développé sur un jeu de données de plus de 300 000 radiographies, réalisées entre Janvier 2011 et Mai 2021 et issues de patients provenant de plus de 60 centres et services de radiologie. Le jeu de données ayant servi au développement a été séparé aléatoirement en 70% pour la phase d'entrainement, 10% pour la phase de validation et 20% pour la phase de test interne. Les prédictions proposées par l'algorithme sont de 3 types en fonction des seuils de fonctionnement pré-établis : « *fract* » (associé à une boîte de localisation en traits pleins), « *no fract* » ou « *doubt fract* » (associé à une boîte de localisation en traits pointillés) (**Figure 1)**

Le jeu de données radiographiques de cette étude n'a été utilisé à aucune étape du développement de l'algorithme et constitue donc un jeu de données de test externe. Il a été analysé en *standalone* par l'algorithme BoneView (Gleamer, Paris, France), en utilisant le seuil de fonctionnement de la version commercialisée, permettant de produire, à l'échelle des fractures, des valeurs de sensibilité, spécificité, VPP, VPN et une courbe FROC (*Free-Response Receiver-Operator Curve*) en comparant la prédiction de l'algorithme sur les radiographies, au *ground truth* scanographique.

<u>Figure 1 :</u> Exemples de radiographies pour lesquelles une fracture du radius (panel supérieur) et du scaphoïde (panel inférieur) ont été correctement identifiées par l'I.A.

ANALYSE DES DONNÉES

Le traitement rétrospectif des données d'imagerie (recueil, dé-identification et export) a été réalisé en septembre 2020 et a fait l'objet d'une déclaration préalable et d'une validation par la Cellule Informatique et Liberté du CHU de Lille (DEC19-279).

L’analyse de la performance individuelle a été réalisée à la fin de la période d’interprétation par les lecteurs. Les performances de l’algorithme et des lecteurs ont été calculées par un membre de Gleamer (J.V.) de façon automatique à partir de la base de données de l’interface de lecture. L’analyse statistique des performances des lecteurs a été effectuée par un auteur indépendant de Gleamer (T.J.) en utilisant le logiciel GraphPad Prism 9 (GraphPad, La Jolla, CA).

La normalité de la distribution des données a été évaluée par méthode de Kolmogorov-Smirnov. Les performances des lecteurs post-IA ont été comparées de façon appariée aux performances pré-IA par ANOVA avec correction *post-hoc* de Šidák pour les comparaisons multiples. Les analyses en sous-groupes selon l’ancienneté des lecteurs ont été réalisées par t-test (apparié ou non selon les sous-groupes étudiés). Le seuil de significativité statistique était défini à p=0.05.

IV - RÉSULTATS

CARACTÉRISTIQUES DU JEU DE DONNÉES

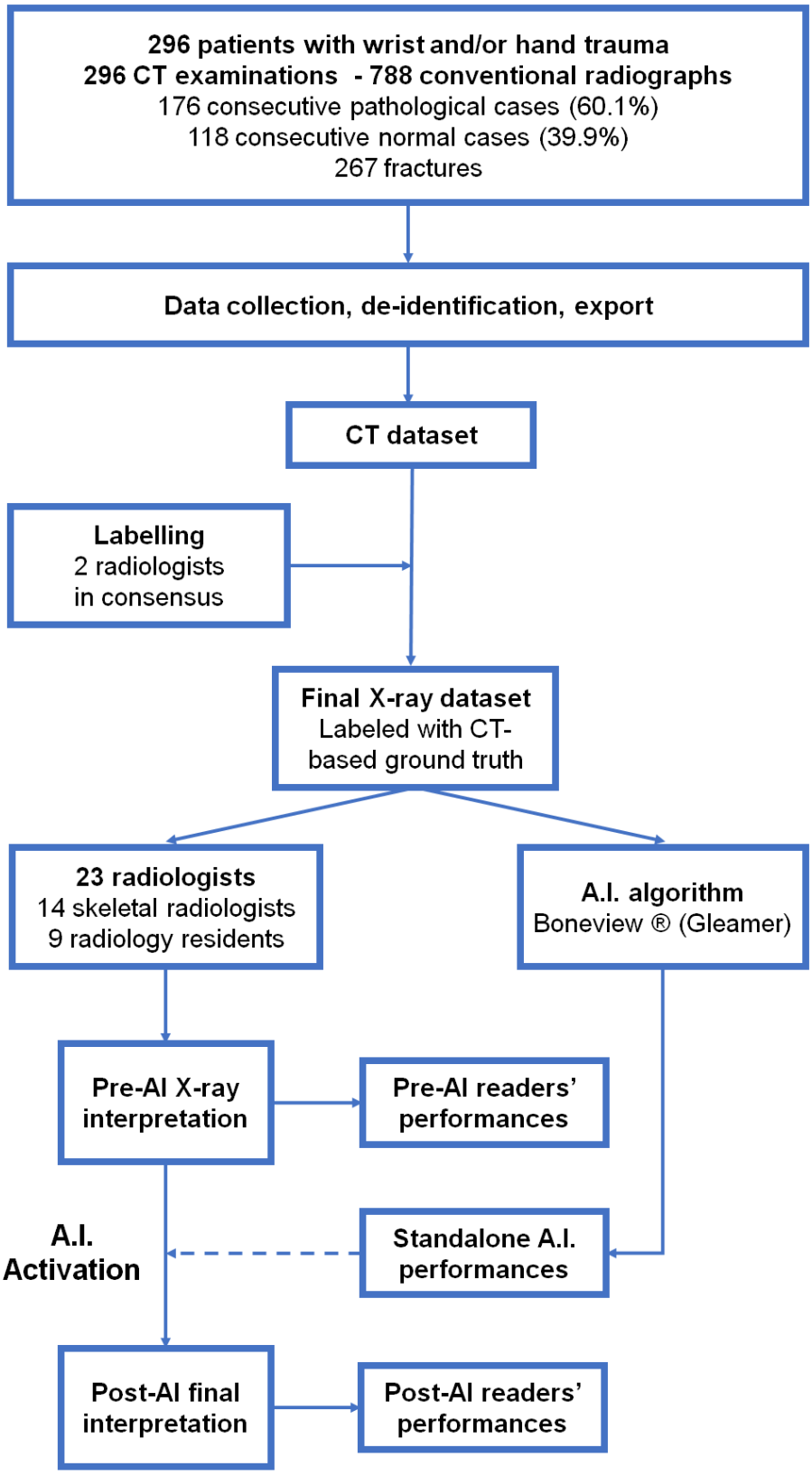
Durant la période de l'étude, 296 patients adultes consécutifs ont bénéficié à la fois de radiographies et d'un scanner pour un traumatisme du poignet et/ou de la main. Le jeu de données final était donc constitué de 296 scanners et des 788 radiographies correspondantes.

Un total de 178 patients (178/296, 60.1%) présentait au moins une anomalie post-traumatique en scanner, pour un total de 267 fractures ou avulsions visibles en scanner (certains patients présentant plusieurs fractures). La prévalence des anomalies post-traumatiques était donc élevée dans ce jeu de données, puisque seulement 118 patients (118/296, 39.9%) n'avaient aucune lésion post-traumatique.

Les fractures intéressaient le radius dans 118 cas (44.2% des fractures), l'ulna dans 43 cas (16.1%), les os du carpe dans 93 cas (34.8%) - dont 35 fractures du scaphoïde et 25 fractures du triquetrum, et les doigts dans 13 cas (4.9%).

Le jeu de données était composé de 52.4% d'hommes et 47.6% de femmes, avec une moyenne d'âge de 41.1 +/- 19.1 ans. Les patients présentant une fracture étaient à 58.4% des hommes et 41.6% des femmes, avec une moyenne d'âge de 42.3 +/- 19.7 ans. Les patients ne présentant pas de fracture étaient à 56.8% des femmes et 43.2% des hommes, avec une moyenne d'âge de 39.1 +/- 18.1 ans.

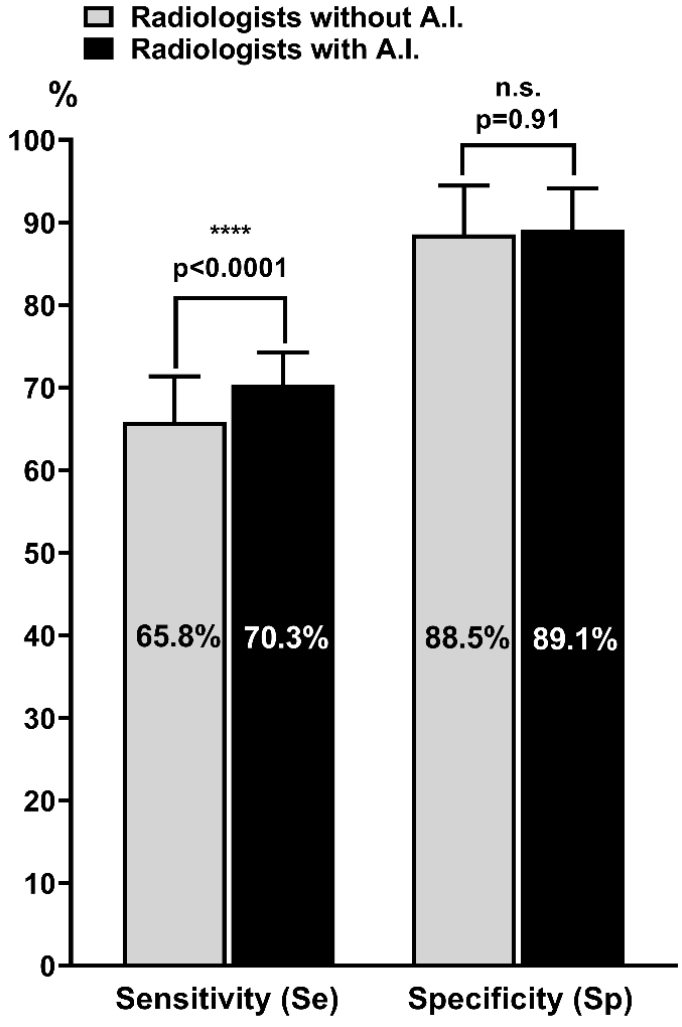
Le plan expérimental de l'étude est présenté en **Figure 2**.



<u>Figure 2 :</u> Plan expérimental de l'étude

MODIFICATIONS DE LA SENSIBILITÉ ET DE LA SPECIFICITÉ DES LECTEURS

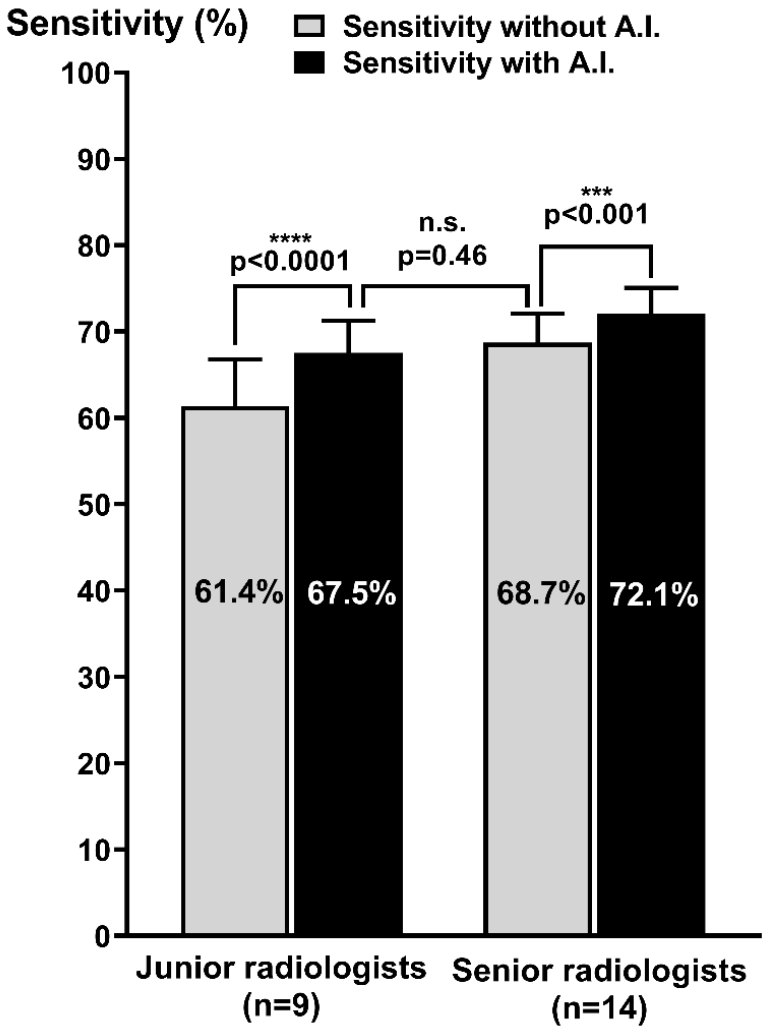
A l'échelle des fractures individuelles, la sensibilité des lecteurs sans l'aide de l'intelligence artificielle était de 65.8% (+/- 5.5%), contre 70.3% (+/- 3.9%) avec l'aide de l'intelligence artificielle, soit un gain de sensibilité moyen de 4.5% (p<0.0001) grâce à l'utilisation de l'algorithme. A l'échelle des patients, la sensibilité passait de 58.2% (+/- 6.2%) à 63.5% (+/- 4.5%) (p<0.0001). A l'inverse, il n'existait pas de différence significative concernant la spécificité des lecteurs, dont la valeur était de 88.5% (+/- 5.9%) sans IA contre 89.1 % (+/- 4.9%) avec IA (p=0.91) (**Figure 3**).



<u>Figure 3:</u> Sensibilité (Se) et Spécificité (Sp) des lecteurs (en %) sans utilisation de l'algorithme d'I.A., puis avec son utilisation.

Le gain de sensibilité par fracture était plus marqué chez les radiologues juniors, avec une augmentation de sensibilité moyenne de 6.1%, passant de 61.4% (+/- 5.4%) à 67.5 % (+/- 3.7%) (p<0.0001) après utilisation de l'algorithme. Les radiologues seniors présentaient un gain de sensibilité moyen de 3.4%, moins marqué mais restant significatif, passant de 68.7 % (+/- 3.4%) à 72.1% (+/- 3.0%) (p<0.001).

De façon intéressante, il n'existait pas de différence significative entre la sensibilité par fracture des radiologues juniors avec IA, estimée à 67.5% (+/- 3.7%), et celle des radiologues seniors sans IA, estimée à 68.7% (+/-3.4%) (p=0.46) (**Figure 4**).

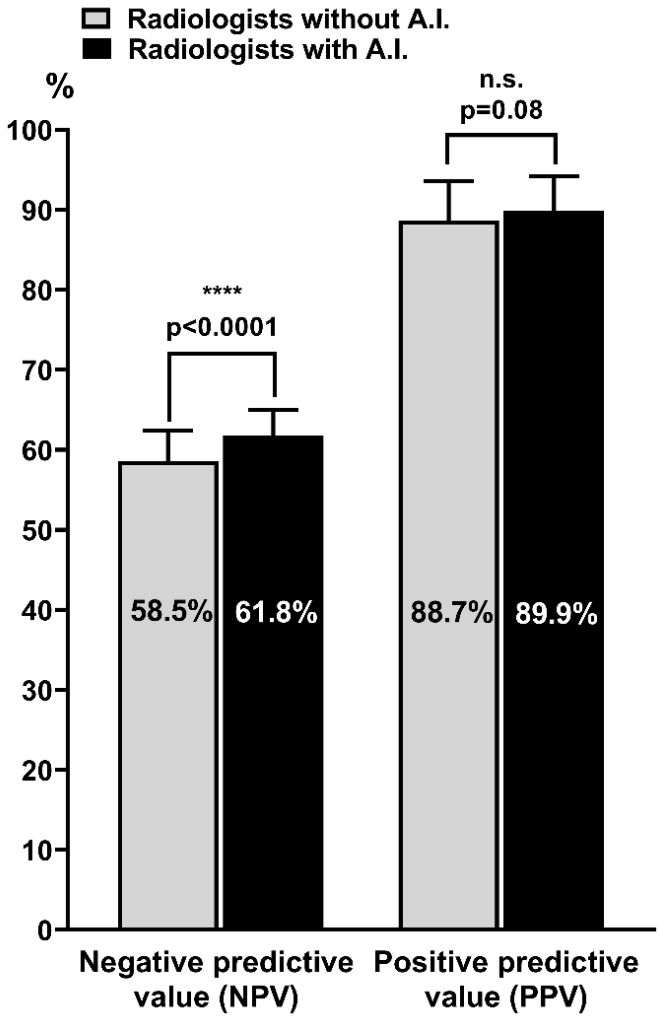


<u>Figure 4:</u> Sensibilité (Se) des lecteurs (en %) sans utilisation de l'algorithme d'I.A., puis avec son utilisation, en fonction du niveau d'expérience (radiologues juniors ou radiologues séniors)

MODIFICATIONS DES VALEURS PRÉDICTIVES POSITIVE ET NÉGATIVE DES LECTEURS

La valeur prédictive négative (VPN) des lecteurs sans l'aide de l'intelligence artificielle était de 58.5% (+/- 3.8%), contre 61.8% (+/- 3.2%) avec l'aide de l'intelligence artificielle, soit un gain moyen de VPN de 3.3% (p<0.001) grâce à l'utilisation de l'algorithme.

A l'inverse, il n'existait pas de différence significative concernant la valeur prédictive positive (VPP) des lecteurs, qui était de 88.7% (+/- 4.9%) sans IA contre 89.9 % (+/- 4.3%) avec IA (p=0.08) (**Figure 5**).



<u>Figure 5 :</u> Valeur prédictive négative (NPV) et positive (PPV) des lecteurs (en %) sans utilisation de l'algorithme d'I.A., puis avec son utilisation.

Le gain de VPN moyen était de 3.8% chez les radiologues juniors, passant de 55.9% (+/- 3.3%) à 59.7 % (+/- 3.1%) (p<0.001) après utilisation de l'algorithme. Les radiologues seniors présentaient un gain de VPN de 2.8%, passant de 60.3 % (+/- 3.2%) à 63.1% (+/- 2.7%) (p<0.05).

Il n'existait pas de différence significative entre la VPN des radiologues juniors avec IA, estimée à 59.7% (+/- 3.1%), et celle des radiologues seniors sans IA, estimée à 60.3% (+/- 3.2%) (p=0.69).

PERFORMANCES DE L'I.A. SEULE (« STANDALONE »)

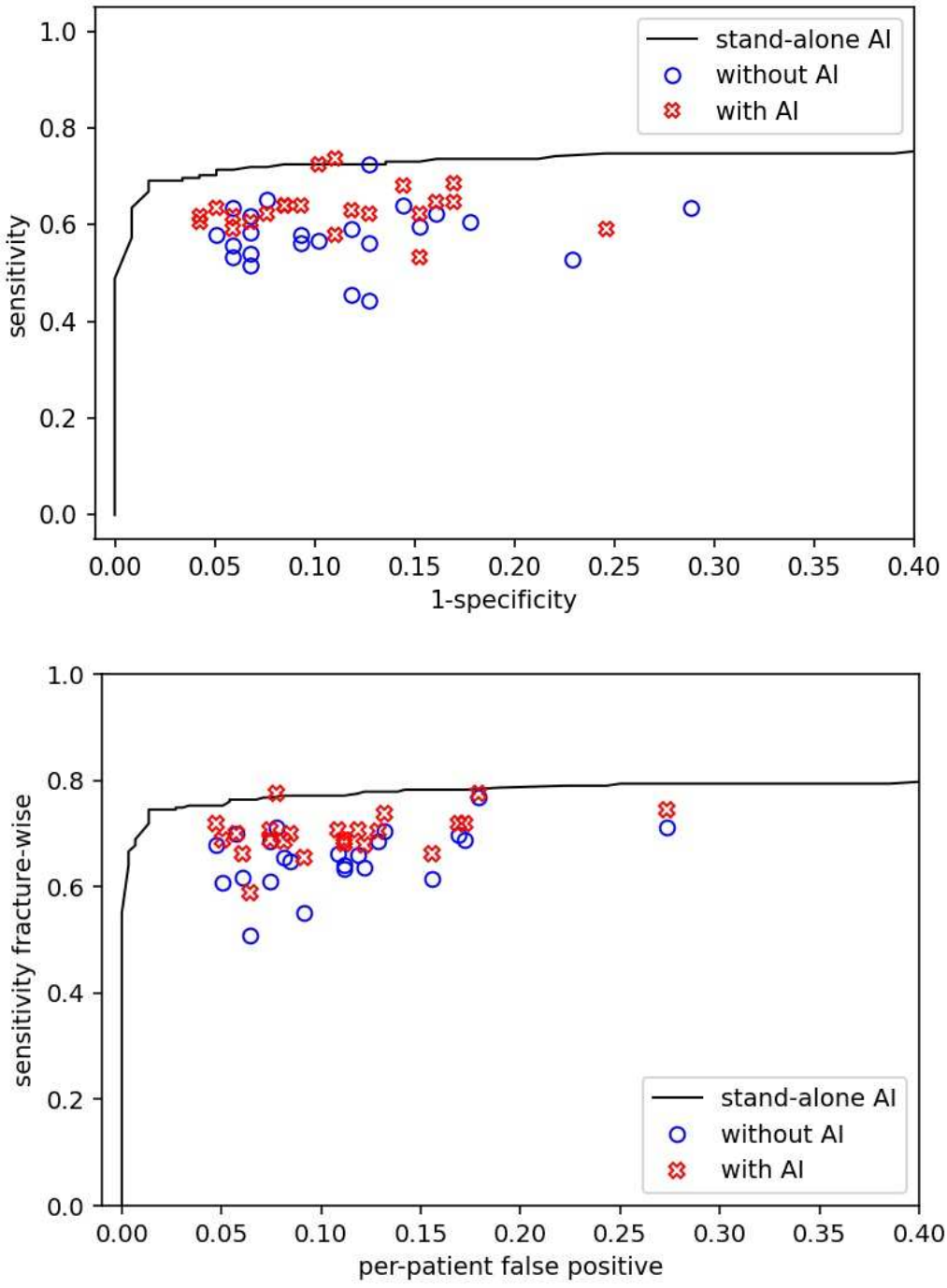
Le jeu de données de radiographies a en parallèle été analysé par l'algorithme, de manière autonome, sans intervention humaine (« *standalone* »). Les prédictions de l'algorithme ont été comparées au *ground truth* scanographique.

Dans ces conditions, la sensibilité de l'algorithme à l'échelle des fractures individuelles était de 77.1%, sa sensibilité à l'échelle des patients de 72.5%, sa spécificité de 89.8%, sa VPN de 68.4% et sa VPP de 91.5%. Ces valeurs étaient proches de celles des 2 lecteurs de l'étude présentant le meilleur indice de Youden (sensibilité + spécificité – 1) sans I.A. (**Tableau 1**).

		Lecteur 1	Lecteur 2	IA seule
Se	**Sans IA**	76.8%	70%	77.1%
	Avec IA	77.5%	70%	
Sp	**Sans IA**	87.3%	94.1%	89.8%
	Avec IA	89.8%	94.9%	
VPN	**Sans IA**	67.8%	63.1%	68.4%
	Avec IA	68.4%	63.3%	
VPP	**Sans IA**	89.6%	94.2%	91.5%
	Avec IA	91.5%	95.0%	

Tableau 1 : Performances de l'algorithme en *standalone*, par rapport aux performances des deux lecteurs ayant les meilleurs indices de Youden sans I.A. (64.1%)

Les courbes ROC (*receiver-operator curve*, à l'échelle des patients) et FROC (*free-response receiver-operator curve*, à l'échelle des fractures) de l'algorithme sur ce jeu de données sont présentées en **Figure 6**.



<u>Figure 6 :</u> Courbes *receiver-operator* (ROC) et *free-response receiver-operator* (FROC) représentant les performances de l'algorithme (trait continu) et celles des lecteurs avec et sans IA. L'aire sous la courbe (AUROC) est respectivement de 0.764 et de 0.790.

PERFORMANCES EN FONCTION DE LA LOCALISATION DE LA FRACTURE

Afin de mieux évaluer l'impact de l'algorithme sur les lecteurs en fonction de la topographie des fractures étudiées, une analyse en sous-groupes a été réalisée. Les résultats sont présentés dans le **Tableau 2**. L'I.A. entraînait une amélioration significative de la sensibilité des lecteurs pour la détection des fractures du radius, du scaphoïde, du triquetrum et des os de la main.

Localisation de le fracture	Sensibilité			
	Lecteurs sans IA	Lecteurs avec IA	Différence	IA seule
Radius (n=118)	**88.0 %** (+/- 4.6%)	**91.8%** (+/-3.0%)	**+ 3.8% (p<0.0001)**	94.9%
Ulna (n=43)	**71.3%** (+/- 8.8%)	**72.4%** (+/-7.8%)	**+ 1.1% (p=0.19)**	93.0%
Scaphoïde (n=35)	**63.9%** (+/-12.3%)	**77.0%** (+/-9.3%)	**+ 13.1% (p<0.0001)**	85.7%
Triquetrum (n=25)	**50.3%** (+/-9.4%)	**54.6%** (+/- 8.8%)	**+ 4.3% (p<0.0001)**	60.0%
Autres os du carpe (n=33)	**10.1%** (+/- 4.4%)	**9.6%** (+/-3.5%)	**- 0.53% (p=0.36)**	15.2%
Os de la main (n=13)	**23.7%** (+/-10.1%)	**28.8%** (+/-7.4%)	**+ 5.1% (p=0.0002)**	30.8%

Tableau 2 : sensibilité des lecteurs et de l'I.A. en fonction de la localisation de la fracture.

V - DISCUSSION

Il s'agit à notre connaissance de la première étude visant à évaluer l'impact d'un algorithme commercial d'IA sur la détection des fractures, qui se base sur un jeu de données radiographiques entièrement validé par un gold standard scanner. Le référentiel scanner a été choisi car il est très probablement un meilleur reflet de la réalité clinique des lésions, puisque les radiographies ont une capacité imparfaite de détection des fractures (5)(8)(9). Une première étude radiographique se basant sur un référentiel scanner a récemment été réalisée en imagerie thoracique (33), et il est probable que d'autres études du même type soient amenées à être publiées dans le futur, car la comparaison aux résultats d'imagerie en coupe permet de mieux apprécier les performances réelles des algorithmes et leurs conséquences potentielles sur la prise en charge ultérieure des patients.

Cette étude a permis de montrer que, même en se référant à un *ground truth* scanner, l'algorithme utilisé (BoneView) était toujours en mesure d'augmenter significativement la sensibilité et la valeur prédictive négative des radiologues pour la détection des fractures du poignet et de la main en radiographie standard, sans altérer leur spécificité ni leur valeur prédictive positive. Le gain moyen de sensibilité était estimé à + 4.5% (à l'échelle des fractures) et + 5.3% (à l'échelle des patients), ce qui était significatif mais avec une valeur légèrement inférieure aux données de la littérature avec le même algorithme (utilisant un *ground truth* radiographique), estimée à +10.4% dans l'étude de Guermazi *et al* (43), et à +8.7% dans l'étude de Duron *et. al.* (44). De plus, notre étude ne retrouvait pas de différence significative en termes de spécificité des lecteurs après utilisation de l'IA, ce qui était concordant avec l'étude de Guermazi *et al.* (43), alors que l'étude de Duron *et al.* (44) retrouvait une augmentation de spécificité de

4.1%. Cependant, ces études s'intéressaient à des fractures toutes zones anatomiques confondues, alors que notre étude s'intéressait spécifiquement au sous-ensemble des fractures du poignet et de la main, ce qui pourrait participer à cette différence.

L'augmentation de la sensibilité et de la valeur prédictive négative sont des données cliniquement pertinentes car elles peuvent se traduire par une diminution des faux négatifs, qui sont à risque de mener à des fractures non diagnostiquées et d'entraîner un pronostic fonctionnel défavorable pour le patient. De plus, l'utilisation de l'IA permettait aux radiologues les moins expérimentés (juniors) d'atteindre des valeurs de sensibilité et de VPN comparables à celles des radiologues plus expérimentés (sans IA). Or, ces lecteurs juniors jouent parfois un rôle majeur dans la première ligne d'interprétation des radiographies aux urgences ; cette augmentation pourrait améliorer la qualité des premières étapes des soins. La spécificité et la VPP des lecteurs, avec et sans IA, étaient bonnes (> 88.5%). Le fait que l'utilisation de l'IA ne dégrade pas significativement ces valeurs élevées est un élément rassurant sur la sécurité des soins. Ces données s'entendent cependant à court terme, car l'impact de l'utilisation à long terme de ce type d'outil sur les performances humaines n'a pas été évalué, et pourrait possiblement mener à une forme de perte de compétence (*deskilling*) (60).

En fonction de la topographie des lésions, la sensibilité rapportée dans la littérature pour la détection des fractures du radius et de l'ulna est estimée aux alentours de 80%, et entre 67% et 80% pour celles du scaphoïde (8). Dans notre étude, les lecteurs sans I.A. avaient une sensibilité par fracture de 88% pour la détection des fractures du radius, 71.3% pour les fractures de l'ulna et 63.9% pour celles du scaphoïde. Les

lecteurs de cette étude avaient donc des performances dans la norme de la littérature, alors même que le jeu de données pouvait être considéré comme difficile. En effet, la prévalence des anomalies post-traumatiques était élevée dans ce jeu de données, puisque seulement 118 patients (118/296, 39.9%) n'avaient aucune lésion post-traumatique.

L'utilisation d'un jeu de données enrichi, ainsi qu'une analyse rétrospective des cas, sont des éléments à risque d'induire une baisse des performances mesurées des lecteurs, par un effet décrit sous le nom de *laboratory effect* (46). De plus, les radiologues étaient en aveugle des renseignements cliniques, ce qui peut potentiellement diminuer leurs performances (52). En effet, il a été démontré que les informations cliniques comme la topographie du traumatisme ou son mécanisme améliorent les performances des radiologues dans la détection d'une fracture, en augmentant notamment leur confiance diagnostique (53). Par exemple, l'utilisation de schémas pour localiser la zone du traumatisme fait passer la sensibilité de détection de fractures subtiles en radiographie de 67% à 73%, et la spécificité de 93% à 94% (54).

Dans l'étude de Guermazi *et al* (43), la sensibilité de détection des radiologues à l'échelle des fractures individuelles était de 76.2% après utilisation de l'IA, et leur spécificité de 95.6%. Dans notre étude, ces valeurs étaient inférieures, respectivement estimées à 70.3% et de 89.1 % après utilisation de l'IA. Cette différence peut provenir des zones anatomiques étudiées dans notre étude (poignet et main), qui sont classiquement plus à risque de fractures non vues. Cependant, cette différence était attendue lors du choix du *ground truth* scanner. En effet, certaines fractures ne sont pas visibles en radiographie y compris par les experts. Ainsi, des radiographies

habituellement considérées comme normales dans les publications sont en réalité pathologiques en utilisant un *ground truth* scanographique, baissant les performances des lecteurs. Néanmoins, ce choix permet un meilleur reflet de la réalité clinique.

De la même façon, les performances de l'algorithme en *standalone* étaient significativement inférieures à celles déjà publiées. En effet, sa sensibilité à l'échelle des patients était de 72.5%, sa spécificité de 89.8% et son AUROC de 0.764, contre respectivement 93%, 100% et 0.94 pour les fractures du poignet et de la main dans l'étude de Guermazi *et al* (43). L'algorithme en *standalone* présentait malgré tout des performances équivalentes ou supérieures à la quasi-totalité des lecteurs, mais cette comparaison n'était pas l'objectif de l'étude. Cependant, la performance des lecteurs et de l'algorithme n'a pas été pondérée par le nombre d'incidences analysées. Une augmentation du nombre d'incidences (clichés obliques, clichés centrés) pourrait modifier ces performances.

Cette diminution franche des performances de l'IA en utilisant un *ground truth* scanographique confirme que ce type d'algorithmes est fortement impacté par les fractures non visibles en radiographie, ce qui semble logique puisqu'elles ne figurent habituellement pas dans ses données d'entrainement. Ces éléments pointent l'intérêt, dans le futur, de disposer de jeux de données labellisés de façon plus robuste (par exemple sur la base d'imageries en coupe ou de données per-opératoires), à la fois pour les phases d'évaluation mais possiblement aussi dès la phase d'entraînement de l'algorithme. La non prise en compte de certaines fractures en radiographie est à risque d'entraîner une stratification cachée (« *Hidden Stratification* ») pouvant affecter la formation du modèle, ses performances et *in fine* ses résultats cliniques (49).

Cette baisse de performance de l'algorithme peut aussi s'expliquer par le caractère « difficile » du jeu de données évalué dans cette étude. En effet, les données étudiées provenaient d'un service d'urgences pour lequel les scanners sont réalisés après un traumatisme du poignet principalement en cas de doute radiologique ou à visée pré-opératoire, ce qui signifie que certaines radiographies étaient potentiellement difficiles et d'autres multi-fracturaires. Or, Raisuddin *et. al.* ont démontré qu'un modèle d'IA entrainé à détecter les fractures du poignet présentait une baisse significative de ses performances diagnostiques sur un jeu de données de validation externe qualifié de difficile, qui avait nécessité un scanner complémentaire afin d'affirmer le diagnostic, avec des AUROCS passant de 0.99 sur le test classique à 0.84 sur le jeu de données difficiles (50).

La question se pose donc de savoir quelle place consacrer à ces algorithmes dans le cadre de la prise en charge des patients aux urgences. Les études récentes tendent à montrer une amélioration des performances des lecteurs humains. Notre étude, bien que basée sur un labelling plus exigeant, conforte ces résultats. Ainsi, l'introduction de ce type d'outil visant à améliorer les performances d'humains dans une filière de prise en charge organisée semble avoir du sens. Néanmoins l'importante baisse d'AUROC en se référant à un *ground truth* scanographique (0.725 dans notre étude contre > 0.90 dans la plupart des publications) confirme que la prédiction d'un algorithme d'IA approche les performances d'experts mais ne doit pas être considérée comme parfaite ni autonome. Ce d'autant que les services d'urgences disposent fréquemment d'un accès au scanner, dont l'AUROC pour les fractures du poignet et de la main est estimée à 0.97 dans la littérature récente (59). De même, des technologies plus

novatrices comme le *cone beam CT* (CBCT) ont des performances similaires, avec une sensibilité dans cette indication estimée entre 87.7% et 90.6% et une spécificité estimée entre 99.2% et 100% (61), avec des dosimétries basses (62). Ainsi, le déploiement d'outils d'IA est une piste très intéressante, mais qui doit s'intégrer dans une filière radiologique organisée et ne doit pas se faire au détriment de prises en charges conventionnelles et d'investissement dans des outils robustes.

Notre article présente toutefois plusieurs limites.

Tout d'abord, le jeu de données était un jeu enrichi, pour limiter le risque de biais par réponse aléatoire des lecteurs (*guessing effect*). Cependant, les lecteurs n'étaient pas informés du niveau d'enrichissement du jeu de données afin de ne pas altérer leur jugement. La prévalence des cas pathologiques étant forte (60%), il est possible que les performances des lecteurs en conditions cliniques (avec une prévalence plus faible) auraient été différentes.

Ensuite, il s'agissait d'une étude monocentrique. Cet élément est moins problématique pour l'évaluation des lecteurs que pour l'évaluation de l'algorithme. En effet, l'hétérogénéité des données analysées est un élément capable d'affecter de façon plus ou moins forte les performances d'un algorithme dans les études de validation externe (45). Nous avons choisi d'utiliser un algorithme marqué, commercialisé et ayant déjà fait l'objet de plusieurs études de validation externe afin de limiter ce risque de fluctuation de performances de l'algorithme.

Par ailleurs, les lecteurs de cette étude étaient uniquement des radiologues. Ce choix a été réalisé pour permettre une homogénéité et mieux évaluer l'impact de l'IA spécifiquement sur cette population de médecins. La cohorte de radiologues de cette étude est d'ailleurs de plus grande taille que dans la plupart des articles sur cette

thématique. Bien que le niveau d'expertise des lecteurs soit variable, de nouvelles études seront nécessaires en incluant d'autres professionnels amenés à interpréter des radiographies dans un contexte post-traumatique, comme des médecins urgentistes ou des chirurgiens orthopédiques.

Enfin, il s'agissait ici d'une étude rétrospective. L'analyse rétrospective des examens d'imagerie par les lecteurs humains est à risque de modifier leurs performances, plutôt en les sous-estimant (46). Concernant les algorithmes d'IA, la relation entre les performances rétrospectives et prospectives est encore incertaine à ce jour, notamment le risque potentiel d'effet-centre, dépendant de la prévalence de la pathologie étudiée (proche ou éloignée des données d'entraînement) et des paramètres techniques des examens analysés.

Des études prospectives seront nécessaires dans un futur proche pour conforter ces résultats et mieux comprendre les modifications de performance des lecteurs et des algorithmes dans des contextes cliniques de soin.

VI - CONCLUSION

L'utilisation d'un algorithme d'IA a permis aux radiologues d'améliorer significativement leur sensibilité et leur valeur prédictive négative pour la détection des fractures du poignet et de la main en radiographie, sans affecter leur spécificité ni leur valeur prédictive positive. Ces données confortent des publications récentes sur cette thématique. L'utilisation d'un *ground truth* scanner comme référence est novateur pour ce type de méthodologie, et est à l'origine de performances significativement inférieures de l'algorithme par rapport à celles rapportées dans la littérature, mais probablement plus proches de la réalité clinique.

RÉFÉRENCES

1. La Statistique annuelle des établissements (SAE) | Direction de la recherche, des études, de l'évaluation et des statistiques [Internet]. [cité 15 mai 2022]. Disponible sur: https://drees.solidarites-sante.gouv.fr/sources-outils-et-enquetes/00-la-statistique-annuelle-des-etablissements-sae

2. Morley C, Unwin M, Peterson GM, Stankovich J, Kinsman L. Emergency department crowding: A systematic review of causes, consequences and solutions. PloS One. 2018;13(8):e0203316.

3. Abimanyi-Ochom J, Watts JJ, Borgström F, Nicholson GC, Shore-Lorenti C, Stuart AL, et al. Changes in quality of life associated with fragility fractures: Australian arm of the International Cost and Utility Related to Osteoporotic Fractures Study (AusICUROS). Osteoporos Int. 2015;26(6):1781-90.

4. de Putter CE, Selles RW, Polinder S, Panneman MJM, Hovius SER, van Beeck EF. Economic Impact of Hand and Wrist Injuries: Health-Care Costs and Productivity Costs in a Population-Based Study. JBJS. 2 mai 2012;94(9):e56.

5. Pinto A, Berritto D, Russo A, Riccitiello F, Caruso M, Paola Belfiore M, et al. Traumatic fractures in adults: missed diagnosis on plain radiographs in the Emergency Department. Acta Bio Medica Atenei Parm. 2018;89(Suppl 1):111-23.

6. Owen RA, Melton LJ, Johnson KA, Ilstrup DM, Riggs BL. Incidence of Colles' fracture in a North American community. Am J Public Health. juin 1982;72(6):605-7.

7. Court-Brown CM, Caesar B. Epidemiology of adult fractures: A review. Injury. 1 août 2006;37(8):691-7.

8. Balci A, Basara I, Çekdemir EY, Tetik F, Aktaş G, Acarer A, et al. Wrist fractures: sensitivity of radiography, prevalence, and patterns in MDCT. Emerg Radiol. juin 2015;22(3):251-6.

9. Welling RD, Jacobson JA, Jamadar DA, Chong S, Caoili EM, Jebson PJL. MDCT and Radiography of Wrist Fractures: Radiographic Sensitivity and Fracture Patterns. Am J Roentgenol. janv 2008;190(1):10-6.

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

11. 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(1):4.

12. Waymel Q, Badr S, Demondion X, Cotten A, Jacques T. Impact of the rise of artificial intelligence in radiology: What do radiologists think? Diagn Interv Imaging. juin 2019;100(6):327-36.

13. Larousse É. intelligence artificielle - LAROUSSE [Internet]. [cité 18 mai 2022]. Disponible sur: https://www.larousse.fr/encyclopedie/divers/intelligence_artificielle/187257

14. Chrusciel J, Le Guillou A, Daoud E, Laplanche D, Steunou S, Clément MC, et al. Making sense of the French public hospital system: a network-based approach to hospital clustering using unsupervised learning methods. BMC Health Serv Res. 17 nov 2021;21(1):1244.

15. Goodfellow I, Bengio Y, Courville A. Deep learning. MIT press; 2016.

16. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design Characteristics of Studies Reporting the Performance of Artificial Intelligence Algorithms for Diagnostic Analysis of Medical Images: Results from Recently Published Papers. Korean J Radiol. mars 2019;20(3):405-10.

17. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. janv 2017;35:303-12.

18. 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. 2 févr 2017;542(7639):115-8.

19. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 13 déc 2016;316(22):2402-10.

20. Sandino CM, Cheng JY, Chen F, Mardani M, Pauly JM, Vasanawala SS. Compressed Sensing: From Research to Clinical Practice with Deep Neural Networks. IEEE Signal Process Mag. janv 2020;37(1):111-27.

21. Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, Suever JD, Geise BD, Patel AA, et al. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. NPJ Digit Med. 4 avr 2018;1:9.

22. Chou PH, Jou THT, Wu HTH, Yao YC, Lin HH, Chang MC, et al. Ground truth generalizability affects performance of the artificial intelligence model in automated vertebral fracture detection on plain lateral radiographs of the spine. Spine J. avr 2022;22(4):511-23.

23. Li YC, Chen HH, Horng-Shing Lu H, Hondar Wu HT, Chang MC, Chou PH. Can a Deep-learning Model for the Automated Detection of Vertebral Fractures Approach the Performance Level of Human Subspecialists? Clin Orthop. 1 juill 2021;479(7):1598-612.

24. Voter AF, Larson ME, Garrett JW, Yu JPJ. Diagnostic Accuracy and Failure Mode Analysis of a Deep Learning Algorithm for the Detection of Cervical Spine Fractures. AJNR Am J Neuroradiol. août 2021;42(8):1550-6.

25. Monchka BA, Kimelman D, Lix LM, Leslie WD. Feasibility of a generalized convolutional neural network for automated identification of vertebral compression fractures: The Manitoba Bone Mineral Density Registry. Bone. sept 2021;150:116017.

26. Chen HY, Hsu BWY, Yin YK, Lin FH, Yang TH, Yang RS, et al. Application of deep learning algorithm to detect and visualize vertebral fractures on plain frontal radiographs. PLoS ONE. 28 janv 2021;16(1):e0245992.

27. Cheng CT, Ho TY, Lee TY, Chang CC, Chou CC, Chen CC, et al. Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol. 2019;29(10):5469-77.

28. Cheng CT, Wang Y, Chen HW, Hsiao PM, Yeh CN, Hsieh CH, et al. A scalable physician-level deep learning algorithm detects universal trauma on pelvic radiographs. Nat Commun. 16 févr 2021;12:1066.

29. Cheng CT, Chen CC, Cheng FJ, Chen HW, Su YS, Yeh CN, et al. A Human-Algorithm Integration System for Hip Fracture Detection on Plain Radiography: System Development and Validation Study. JMIR Med Inform. 27 nov 2020;8(11):e19416.

30. Sato Y, Takegami Y, Asamoto T, Ono Y, Hidetoshi T, Goto R, et al. Artificial intelligence improves the accuracy of residents in the diagnosis of hip fractures: a multicenter study. BMC Musculoskelet Disord. 3 mai 2021;22:407.

31. Bae J, Yu S, Oh J, Kim TH, Chung JH, Byun H, et al. External Validation of Deep Learning Algorithm for Detecting and Visualizing Femoral Neck Fracture Including Displaced and Non-displaced Fracture on Plain X-ray. J Digit Imaging. 1 oct 2021;34(5):1099-109.

32. Hongbiao S, Shaochun X, Xiang W, YuRun T, Yang L, Mingzi Z, et al. Comparison and verification of two deep learning models for the detection of chest CT rib fractures. Acta Radiol Stockh Swed 1987. 18 mars 2022;2841851221083519.

33. Gipson J, Tang V, Seah J, Kavnoudias H, Zia A, Lee R, et al. Diagnostic accuracy of a commercially available deep-learning algorithm in supine chest radiographs following trauma. Br J Radiol. 24 mars 2022;20210979.

34. Zhou QQ, Hu ZC, Tang W, Xia ZY, Wang J, Zhang R, et al. Precise anatomical localization and classification of rib fractures on CT using a convolutional neural network. Clin Imaging. janv 2022;81:24-32.

35. Olczak J, Fahlberg N, Maki A, Razavian AS, Jilert A, Stark A, et al. Artificial intelligence for analyzing orthopedic trauma radiographs. Acta Orthop. nov 2017;88(6):581-6.

36. Tobler P, Cyriac J, Kovacs BK, Hofmann V, Sexauer R, Paciolla F, et al. AI-based detection and classification of distal radius fractures using low-effort data labeling: evaluation of applicability and effect of training set size. Eur Radiol. 2021;31(9):6816-24.

37. Blüthgen C, Becker AS, Martini IV de, Meier A, Martini K, Frauenfelder T. Detection and localization of distal radius fractures: Deep learning system versus radiologists. Eur J Radiol [Internet]. 1 mai 2020 [cité 23 mars 2022];126. Disponible sur: https://www.ejradiology.com/article/S0720-048X(20)30114-5/fulltext

38. 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.

39. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional Neural Networks for Automated Fracture Detection and Localization on Wrist Radiographs. Radiol Artif Intell. 30 janv 2019;1(1):e180001.

40. 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. août 2019;90(4):394-400.

41. AI Central [Internet]. [cité 11 mai 2022]. Disponible sur: https://aicentral.acrdsi.org/

42. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 1 mai 2018;73(5):439-45.

43. 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. 21 déc 2021;210937.

44. 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.

45. Yu AC, Mohajer B, Eng J. External Validation of Deep Learning Algorithms for Radiologic Diagnosis: A Systematic Review. Radiol Artif Intell [Internet]. 4 mai 2022 [cité 11 mai 2022]; Disponible sur: https://pubs.rsna.org/doi/epdf/10.1148/ryai.210064

46. Gur D, Bandos AI, Cohen CS, Hakim CM, Hardesty LA, Ganott MA, et al. The « laboratory » effect: comparing radiologists' performance and variability during prospective clinical and laboratory mammography interpretations. Radiology. oct 2008;249(1):47-53.

47. Moons KGM, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart Br Card Soc. mai 2012;98(9):691-8.

48. Oliveira e Carmo L, van den Merkhof A, Olczak J, Gordon M, Jutte PC, Jaarsma RL, et al. An increasing number of convolutional neural networks for fracture recognition and classification in orthopaedics. Bone Jt Open. 20 oct 2021;2(10):879-85.

49. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proc ACM Conf Health Inference Learn. avr 2020;2020:151-9.

50. Raisuddin AM, Vaattovaara E, Nevalainen M, Nikki M, Järvenpää E, Makkonen K, et al. Critical evaluation of deep neural networks for wrist fracture detection. Sci Rep. 16 mars 2021;11:6006.

51. Ozkaya E, Topal FE, Bulut T, Gursoy M, Ozuysal M, Karakaya Z. Evaluation of an artificial intelligence system for diagnosing scaphoid fracture on direct radiography. Eur J Trauma Emerg Surg Off Publ Eur Trauma Soc. févr 2022;48(1):585-92.

52. Castillo C, Steffens T, Sim L, Caffery L. The effect of clinical information on radiology reporting: A systematic review. J Med Radiat Sci. mars 2021;68(1):60-74.

53. Berbaum KS, el-Khoury GY, Franken EA, Kathol M, Montgomery WJ, Hesson W. Impact of clinical history on fracture detection with radiography. Radiology. août 1988;168(2):507-11.

54. Sarwar A, Wu JS, Kung J, Brook A, Lee KS, Gauguet JM, et al. Graphic representation of clinical symptoms: a tool for improving detection of subtle fractures on foot radiographs. AJR Am J Roentgenol. oct 2014;203(4):W429-433.

55. Lim SS, Phan TD, Law M, Goh GS, Moriarty HK, Lukies MW, et al. Non-radiologist perception of the use of artificial intelligence (AI) in diagnostic medical imaging reports. J Med Imaging Radiat Oncol. 21 févr 2022;

56. Yakar D, Ongena YP, Kwee TC, Haan M. Do People Favor Artificial Intelligence Over Physicians? A Survey Among the General Population and Their View on Artificial Intelligence in Medicine. Value Health. mars 2022;25(3):374-81.

57. Lennartz S, Dratsch T, Zopfs D, Persigehl T, Maintz D, Hokamp NG, et al. Use and Control of Artificial Intelligence in Patients Across the Medical Workflow: Single-Center Questionnaire Study of Patient Perspectives. J Med Internet Res. 17 févr 2021;23(2):e24221.

58. Krastman P, Mathijssen NM, Bierma-Zeinstra SMA, Kraan G, Runhaar J. Diagnostic accuracy of history taking, physical examination and imaging for phalangeal, metacarpal and carpal fractures: a systematic review update. BMC Musculoskelet Disord. 7 janv 2020;21(1):12.

59. Brink M, Steenbakkers A, Holla M, de Rooy J, Cornelisse S, Edwards MJ, et al. Single-shot CT after wrist trauma: impact on detection accuracy and treatment of fractures. Skeletal Radiol. juin 2019;48(6):949-57.

60. Levy J, Jotkowitz A, Chowers I. Deskilling in ophthalmology is the inevitable controllable? Eye Lond Engl. mars 2019;33(3):347-8.

61. Fitzpatrick E, Sharma V, Rojoa D, Raheman F, Singh H. The use of cone-beam computed tomography (CBCT) in radiocarpal fractures: a diagnostic test accuracy meta-analysis. Skeletal Radiol. mai 2022;51(5):923-34.

62. Jacques T, Morel V, Dartus J, Badr S, Demondion X, Cotten A. Impact of introducing extremity cone-beam CT in an emergency radiology department: A population-based study. Orthop Traumatol Surg Res OTSR. avr 2021;107(2):102834.

AUTEUR : Nom : CARDOT **Prénom : Nicolas**

Date de soutenance : 21 Juin 2022

Titre de la thèse : Évaluation de l'impact d'un algorithme d'intelligence artificielle sur la détection par les radiologues des fractures du poignet et de la main en radiographie

Thèse - Médecine - Lille 2022

Cadre de classement : *Radiodiagnostic et Imagerie Médicale*

DES + FST/option : *DES de Radiodiagnostic et Imagerie Médicale*

Mots-clés : Radiographie, Main et poignet, Intelligence artificielle, Validation externe

Résumé :

INTRODUCTION : Les fractures non détectées sont la cause principale d'erreurs médicales aux urgences, en particulier les fractures du poignet de de la main. Des algorithmes d'intelligence artificielle (I.A.) spécialisés dans l'aide à la détection de fractures en radiographie standard ont été développés ces dernières années et se déploient en pratique clinique. Néanmoins, la référence de ces algorithmes est uniquement radiographique, et peut donc sous-estimer les fractures par rapport à un ground truth scanographique. L'objectif de cette étude était d'évaluer les performances de radiologues sans puis après l'aide d'un algorithme d'I.A. (BoneView®) dans la détection des fractures du poignet et de la main en radiographie standard, en utilisant le scanner comme *ground truth*.

MATÉRIEL ET MÉTHODES : Un jeu de données composé de radiographies de la main et du poignet ainsi que des scanners concomitants, réalisés dans un contexte post-traumatique aux urgences d'un centre hospitalier et universitaire, a été constitué de façon rétrospective. Toutes les radiographies ont été annotées par deux radiologues, en consensus, en fonction des anomalies post-traumatiques visibles en scanner. Les données étaient composées de 296 patients, 118 ne présentant aucune fracture (39.9%) et 178 avec au moins une fracture (60.1%) pour un total de 267 fractures visibles en scanner. Vingt-trois radiologues avec des niveaux d'expertise différente (14 radiologues seniors, 9 internes de radiologie) ont été inclus pour analyser rétrospectivement l'ensemble des radiographies, sans puis avec l'aide de l'I.A., en aveugle des résultats du scanner.

RÉSULTATS : Sur les données radiographiques, en se basant sur le ground truth scanner, les performances de l'IA pour la détection de fracture étaient les suivantes : 72.5% (sensibilité), 89.8% (spécificité), 68.4% (valeur prédictive négative - VPN), 91.5% (valeur prédictive positive - VPP) et 0.764 (AUROC). L'utilisation de l'algorithme a permis aux radiologues d'améliorer leur sensibilité, passant de 65.8% à 70.3% à l'échelle des fractures ($p<0.0001$) et de 58.2% à 63.5% à l'échelle des patients ($p<0.0001$), ainsi que leur VPN, passant de 58.5% à 61.8% ($p<0.0001$), indépendamment de leur niveau d'expertise. La spécificité des lecteurs n'était pas significativement affectée, passant de 88.5% à 89.1% ($p=0.91$) de même que leur VPP, de 88.7% à 89.9% ($p=0.08$).

CONCLUSION : L'utilisation d'un algorithme d'I.A. a permis aux radiologues d'améliorer significativement leur sensibilité et leur valeur prédictive négative pour la détection des fractures du poignet et de la main en radiographie, sans affecter ni leur spécificité ni leur valeur prédictive positive. Ces données confortent des publications récentes sur cette thématique. L'utilisation d'un ground truth scanner comme référence est novateur pour ce type de méthodologie, et est à l'origine de performances significativement inférieures de l'algorithme par rapport à celles rapportées dans la littérature, mais probablement plus proches de la réalité clinique.

Composition du Jury :

Présidente : Madame le Professeur Anne COTTEN
Assesseurs : Monsieur le Professeur Christophe CHANTELOT, Monsieur le Professeur Xavier DEMONDION, Monsieur le Professeur Eric WIEL
Directeur de thèse : Monsieur le Docteur Thibaut JACQUES