Méthode d'évaluation des reconstructions 3D de colonne vertébrale, issues d'images radiographiques bi-planaires

par

Magali BONHOMME

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
EN GÉNIE CONCENTRATION TECHNOLOGIES DE LA SANTÉ
M.SC. A

MONTRÉAL, LE 6 AOÛT 2021

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE :

M. Jacques A. De Guise, directeur de mémoire
Département de génie des systèmes à l'École de technologie supérieure

M. Carlos Vázquez, codirecteur de mémoire
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Simon Drouin président du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

Mme Nicola Hagermeister, membre du jury
Département de génie des systèmes à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 29 JUILLET 2021

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je remercie en tout premier mon directeur de recherche, Jacques, qui m'a donné sa pleine confiance et l'autonomie nécessaire pour ce type de projet. Tu m'as insufflé la juste dose de leadership et de persévérance pour aller jusqu'au bout.

Je remercie mon co-directeur, Carlos, pour sa bienveillance, ses encouragements et sa confiance à chacune des nombreuses réunions de ce projet, tu m'as permis de partir dans la bonne direction et de ne pas me perdre ! Merci également à Thierry, sans qui ce projet ne serait tout simplement jamais parti. Malgré la pandémie tu as réussi à garder un lien fort avec les étudiants, et m'a permis aussi de rester dans la ligne tout au long de ce projet. Je tiens à remercier aussi Nicola, qui m'a donné de précieux conseils tout au long de ma maîtrise, et qui a participé grandement à la confiance que j'ai pu donner à mon travail.

Milles merci à mes merveilleux collègues au LIO, j'aurais aimé passer plus de temps au bureau avec vous cette dernière année. Merci Yoyo, mon acolyte du café et des jolis dessins de motivation ; Merci Amiel, pour nos ronchonnades communes, nos rigolades et nos échanges de photos de chats pour oublier le travail ; Merci Sarra, à ton amitié, à nos conversations tard le soir quand tout le monde est parti ; Merci Anaïde, je loue ta patience et ton calme olympique, ça aurait presque donné envie de faire un doctorat (presque) ; Un gros merci à toi Romain, notre premier prix restera affiché fièrement longtemps sur mon CV. Merci à tous les autres, Marie pour m'avoir fait confiance dans cette histoire de maîtrise, dans cette histoire de webinaire, et je l'espère dans d'autres histoires professionnelles palpitantes. Merci Dévine, tu sais comme le chemin a été long et pierreux, merci de m'avoir continuellement soutenue, coachée et encouragée.

Un énorme merci également à la team EOS ; Christine je te remercie de m'avoir épaulée et coachée, Pierre pour ton aide constante pendant tout ce projet, Benjamin pour ton soutien et ta bienveillance! Le plus gros merci du monde à Emmanuelle, sans toi ce projet, je n'y serais pas arrivée. Merci pour ton soutien sans faille, ton intérêt et ta confiance. Merci aussi à Manuela qui depuis la France m'a guidée sur de nombreux aspects de ce projet! Merci à Jay et Gab pour votre soutien et votre bonne humeur au 3DS. Merci Nasr, Paul, et Lukas.

Un énorme merci à Chacha et surtout à Maman, qui m'a encouragée depuis le tout premier moment où je suis montée dans le train. Merci à ma famille du Canada, Barbiche, Sese, vous y croyiez plus et pourtant! Merci Julien, mister rougail-cookie-boardgames. Merci à ma famille, Pierre-Louis et Jimmy.

Méthode d'évaluation des reconstructions 3d de colonne vertébrale, issues d'images radiographiques bi-planaires

Magali BONHOMME

RÉSUMÉ

Les méthodes d'analyses d'images médicales permettent de réaliser des mesures sur des structures biologiques, directement ou par l'intermédiaire de détection des structures (segmentation et extraction d'objets 3D pour visualisation). Quel que soit le domaine d'application ou la modalité d'imagerie, de nombreux algorithmes automatiques sont développés pour améliorer la rapidité, la fiabilité et la précision de ces mesures. Il manque par ailleurs une standardisation des méthodes d'évaluations de cette fiabilité et de cette précision, encore plus lorsqu'il n'y a pas d'étalon or auquel se comparer. De plus, très souvent les objectifs cliniques derrière la tâche d'analyse d'image ne sont pas pris en compte pour l'évaluation de la performance. Lors de reconstructions 3D de colonne vertébrale pour l'évaluation de la gravité des scolioses par exemple, une erreur sur la mesure déterminant la décision de chirurgie du patient aura beaucoup plus d'impacts sur le patient qu'une erreur sur une mesure secondaire qui ne joue aucun rôle sur la décision chirurgicale. Dans les deux cas il existe des impacts pour le patient, mais avec une gravité différente qu'il faut être capable de discriminer pour mieux caractériser la performance de la méthode d'évaluation.

L'objectif de ce travail est de participer à améliorer les méthodes d'évaluations en l'absence d'étalon or en proposant une méthodologie de création de référence qui s'appuie sur l'expérience de terrain, et une méthodologie d'évaluation qui prenne en compte des critères cliniques qui permettent de nuancer les performances atteintes par une méthode automatique.
La méthode d'évaluation propose d'inclure une méthodologie de production de référence à l'aide d'experts qui permet d'obtenir des intervalles de confiance de mesures plus fiables, en incluant des discussions entre les experts. La méthode d'évaluation inclut ces intervalles de confiance ainsi que de nouveaux critères cliniques inspirés des sociétés savantes liées à notre domaine d'application, l'analyse de la scoliose, et à notre modalité d'imagerie, les reconstructions 3D issues de radiographies bi-planaires avec le système EOS. La méthodologie mise en place est généralisable à d'autres domaines et d'autres modalités d'imagerie où les mêmes problématiques sont rencontrées.

Cette approche a été par la suite transférée à l'entreprise partenaire pour pouvoir conduire des évaluations sur les algorithmes automatiques de reconstructions 3D de colonne vertébrale pour les patients pédiatriques et les patients adultes présentant des déformations de la colonne.

Mots-clés : réalité terrain, intervalles de confiance, méthode d'évaluation, reconstructions 3D de colonne vertébrale

Evaluation method of 3d spine reconstructions obtained from biplanar radiographs

Magali BONHOMME

ABSTRACT

Medical image analysis methods allow measurements to be made on biological structures, either by direct measurements or through structures detection (segmentation and extraction of 3D objects for visualization). Regardless of the field of application or the imaging modality, many automatic algorithms are developed to improve the reliability and accuracy of these measurements. On the other hand, there is a lack of standardization of methods for assessing this reliability and accuracy, even more so when there is no gold standard reference to compare. In addition, the clinical goals behind the image analysis task are often not considered for performance evaluation, and therefore there is no hierarchical information between errors. For example, for 3D reconstructions of the scoliotic spine, an error on a measure which determines the surgical decision for the patient would have more impact than an error on a secondary measure that does not interfere with the surgical decision.

The objective of this work is to participate in improving the evaluation methods in the absence of a gold standard by proposing a reference construction methodology based on field experience, and an evaluation methodology that considers the clinical criteria that allow to qualify the performance achieved by an automatic method. The proposed evaluation approach allows for the inclusion of a methodology for reference with the help of experts that allows for more reliable measurement confidence intervals on images, including discussions between experts. The evaluation method includes these confidence intervals as well as new clinical criteria inspired by dedicated research and medical societies related to our field of application, the analysis of scoliosis, and our imaging modality, 3D reconstructions from two-plane X-rays with the EOS system. The methodology put in place is generalizable to other areas and other imaging modalities where the same problems would be encountered.

This method was subsequently transferred to the partner company to be able to conduct evaluations on automatic 3D spinal reconstruction algorithms for pediatric patients and adult patients with spinal deformities.

Keywords: ground truth, confidence intervals, evaluation method, 3D spinal reconstructions

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

2D	Bidimensionnel, 2 dimensions
3D	Tridimensionnel, 3 dimensions
AP	Antéro-Postérieur, vue frontale
CT	Tomodensitométrie
EOS	Système d'acquisition de radiographie biplane
IRM	Imagerie par Résonance Magnétique
ISO	Organisation internationale de standardisation
LAT	Latéral, vue latérale
LIO	Laboratoire de recherche en Imagerie et Orthopédie
RMSSD	Moyenne quadratique des écarts types
SD	Écart-type
SOSORT	Société Scientifique Internationale sur les traitements orthopédiques et conservateurs de la Scoliose
SRS	Société de recherche sur la Scoliose
STAPLE	Estimation simultanée de la vérité terrain et du niveau de performance

INTRODUCTION

Le travail présenté dans ce mémoire s'inscrit dans le contexte de l'évaluation des méthodes d'analyse d'images biomédicales, qui regroupent des tâches allant de mesures simples effectuées dans les images à des segmentations fines de structures dans les images. Ces tâches sont essentielles pour la prise en charge médicale de nombreuses pathologies, pour poser le diagnostic, pour surveiller l'évolution de la maladie ou pour planifier son traitement. Ces tâches font appel à des méthodes de plus en plus automatisées et l'évaluation de leurs performances repose sur trois axes : définir une référence, quantifier les écarts à la référence, et enfin qualifier les erreurs observées. Il n'y a cependant pas de standards quant à la manière d'évaluer ces nouvelles méthodes.

Une méta-analyse sur les approches d'évaluations employées dans le cadre de 150 défis d'analyses d'images, conduits dans les grandes conférences scientifiques internationales jusqu'en 2016, Maier-Hein *et al.* (2018) a mis en évidence le manque de pratiques communes. L'objectif de ces défis est de confronter diverses méthodes d'analyses d'images sur leurs performances pour une tâche donnée, afin de stimuler la recherche scientifique, accroitre les connaissances scientifiques, et permettre aux différentes équipes de recherche dans le monde de positionner leurs méthodes d'analyses les unes par rapport aux autres. Il est ressorti de cette méta-analyse que les méthodes de définition des références d'évaluations étaient très hétérogènes, avec un manque de guides de bonne pratique ou de recommandations. En effet, dans 66% des défis analysés, aucune description de la construction de la référence n'est incluse. Dans 62% des défis de segmentation en particulier, le nombre d'observateurs ayant participé à la construction de la référence n'est pas précisé, et dans 45% de tous les défis il n'est jamais fait mention de comment les avis experts ont été combinés pour construire la référence d'évaluation. Concernant les quantifications des écarts à la référence, les métriques utilisées n'étaient justifiées que dans 23% des défis. De plus, Maier-Hein *et al.* (2018) a identifié un manque de robustesse dans certains classements de méthodes, avec des changements de métriques qui modifient radicalement ces résultats. Il existe pourtant dans la littérature des discussions permettant d'identifier les plus pertinentes selon l'objectif d'évaluation (Taha & Hanbury, 2015). Enfin, concernant la qualification des

erreurs observées, Maier-Hein *et al.* (2018) ne semblent pas avoir investigué cet aspect des évaluations de méthodes, mais le questionnaire final soumis à 295 participants des défis permet cependant d'identifier un besoin dans la communauté scientifique de fournir davantage de documentations sur l'organisation des défis (47% des répondants sur la partie documentation des défis du questionnaire). De même, il existe un besoin de réfléchir sur la pertinence des méthodes de classement (11% des répondants sur la partie évaluation), ou encore de réfléchir au peu d'attention accordée aux besoins cliniques (3% des répondants sur la partie évaluation). Globalement, les principales problématiques soulevées par les participants étaient le manque de représentativité des données (33% des répondants pour la partie données du questionnaire), la qualité des données de référence (33% des répondants pour la partie annotations du questionnaire), l'exhaustivité et la transparence de la documentation des défis (47% des répondants pour la partie documentation du questionnaire), et enfin le choix des métriques et le manque de standard d'évaluations (respectivement 20% et 19% des répondants pour la partie évaluation du questionnaire).

Le travail présenté dans ce mémoire est centré sur l'évaluation de méthodes de reconstruction 3D de colonne vertébrale à partir de radiographies bi planaires. La compagnie EOS Imaging est un partenaire industriel du laboratoire de recherche en Imagerie et Orthopédie de Montréal, et cette collaboration consiste à développer, concevoir et améliorer les méthodes de reconstruction 3D.

La reconstruction 3D de colonne vertébrale présente l'inconvénient de ne pas disposer d'étalon or pour le rachis entier auquel se comparer pour des études de précision. Un étalon or est par définition la meilleure référence possible pour une méthode d'analyse, car il contient la plus grande précision vérifiée et validée pour une analyse donnée (Cardoso et al., 2014), et le but serait de s'en rapprocher le plus avec une nouvelle méthode d'analyse que l'on souhaite évaluer. Pour les mesures cliniques issues des reconstructions 3D et utilisées pour la prise en charge du patient, des études de reproductibilité ont pu être conduites en dépit du manque d'étalon or pour la position debout (Ilharreborde *et al.*, 2016; Somoskeöy *et al.*, 2012). Pour le développement de nouvelles méthodes de reconstructions 3D, il paraît de plus nécessaire de pouvoir inclure dans l'évaluation de performance des aspects cliniques

afin de mieux guider les développements et les améliorations. On constate donc un grand besoin pour une méthode d'évaluation qui combine une référence de qualité ainsi que des critères d'évaluation qui soient pertinents d'un point de vue clinique.

La revue de littérature au chapitre 1 de ce mémoire s'attache à identifier parmi les trois axes d'évaluation précédemment cités (définition de référence, quantification et qualification des erreurs) les pratiques courantes de la littérature, tout domaine d'imagerie considéré, afin d'obtenir une vision claire de ce qui est utilisé aujourd'hui. Après l'exposition de notre problématique générale et de nos objectifs au chapitre 2, nous proposons au chapitre 3 une méthodologie de construction de référence pour les mesures cliniques issues des reconstructions 3D de colonne vertébrale. Le chapitre 4 consiste à mettre en place une méthode d'évaluation qui prenne en compte notre référence et les aspects cliniques liés à la reconstruction 3D du rachis. Enfin, au chapitre 5 nous analyserons les liens entre les modèles 3D du rachis et les mesures cliniques issues automatiquement de ces modèles 3D, afin de dégager des perspectives d'évolution à notre outil d'évaluation.

CHAPITRE 1

REVUE DE LA LITTÉRATURE

1.1 Introduction

Plusieurs approches existent pour analyser une image médicale, qui sont de plus en plus automatisées avec les nouvelles techniques d'intelligence artificielle. Cette automatisation graduelle pose le défi de pouvoir évaluer la qualité de ces analyses d'images par rapport à une référence qui doit être dans l'idéal considérée comme un étalon or, soit une vérité absolue et vérifiée qu'il faudrait approcher le plus possible (Heimann *et al.*, 2009; Taha & Hanbury, 2015). Des défis se posent également quant à la quantification des erreurs d'une méthode d'analyse ainsi qu'à leurs qualifications (comment donner une signification contextuelle à ces erreurs).

Proposer un cadre d'évaluation objectif pour des méthodes d'analyse d'images cliniques serait une étape critique dans la validation et l'applicabilité clinique d'un algorithme (Udupa *et al.*, 2006). Udupa *et al.* (2006) a d'ailleurs identifié plusieurs problématiques dans les méthodes d'évaluation de performance d'algorithmes, parmi lesquelles la difficulté de déterminer une référence, la pauvre définition des métriques de performance, et le fait que les algorithmes ne soient pas comparés à d'autres algorithmes réalisant la même analyse d'image.

Pour notre revue de la littérature, et dans la suite des constats de Maier-Hein *et al.* (2018) sur l'hétérogénéité des pratiques d'évaluations en analyse d'image, nous nous intéresserons aux différentes approches de construction de référence, aux différentes métriques de quantification d'erreurs utilisées, ainsi qu'aux différents critères utilisés pour qualifier les erreurs d'analyse. Ce travail rejoint également le cadre idéal d'évaluation de Udupa *et al.* (2006), dans lequel la construction d'une référence de qualité et la spécification de métriques et de critères de performance sont des composantes importantes.

Nous souhaitons mettre en place une nouvelle méthode d'évaluation pour la reconstruction 3D de colonne vertébrale avec le système EOS de radiographie, nous définirons donc en premier lieu les termes liés à ce domaine d'application.

1.2 La colonne vertébrale

1.2.1 Anatomie

Le rachis, ou colonne vertébrale, est une structure osseuse constituée de segments mobiles s'étendant de la base du crâne au bassin. Le rachis a une fonction de soutien de l'ensemble du squelette, ainsi qu'une fonction de protection du système nerveux central (moelle épinière). Le rachis sain sans anomalies est composé de 7 vertèbres cervicales (C1 à C7), 12 vertèbres thoraciques (T1 à T12), 5 vertèbres lombaires (L1 à L5), du sacrum, et du coccyx. Les côtes formant la cage thoracique sont articulées avec les 12 vertèbres thoraciques. Droit dans le plan frontal, le rachis a des courbures naturelles dans le plan sagittal, qui sont physiologiques

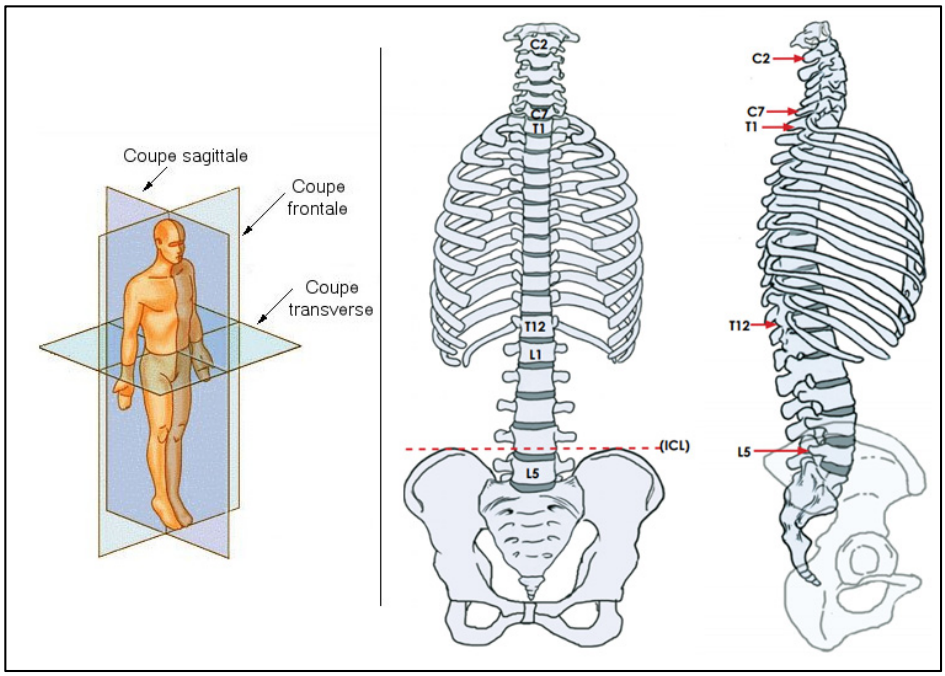


Figure 1.1 Plans anatomiques de référence (à gauche) et anatomie de la colonne vertébrale (à droite)
Adaptée de M. F. O'Brien & Spinal Deformity Study Group (2004, p. 8)

et participent à l'équilibre du corps (voir figure 1.1) (M. F. O'Brien & Spinal Deformity Study Group, 2004).

La scoliose est une déformation tri dimensionnelle de la colonne vertébrale. Pour les adolescents, on parle généralement de scoliose adolescente idiopathique, et pour les adultes de scoliose adulte dégénérative (Bettany-Saltikov *et al.*, 2017) (voir figure 1.2). Cette pathologie peut être évaluée par imagerie, notamment avec la radiographie conventionnelle, le scanner (aussi appelé CT pour *computerized tomography* ou tomodensitométrie), ou le système EOS de radiographie. L'évaluation des déformations à l'aide d'imagerie consiste à réaliser des mesures cliniques, qui seront définies plus loin dans ce chapitre.

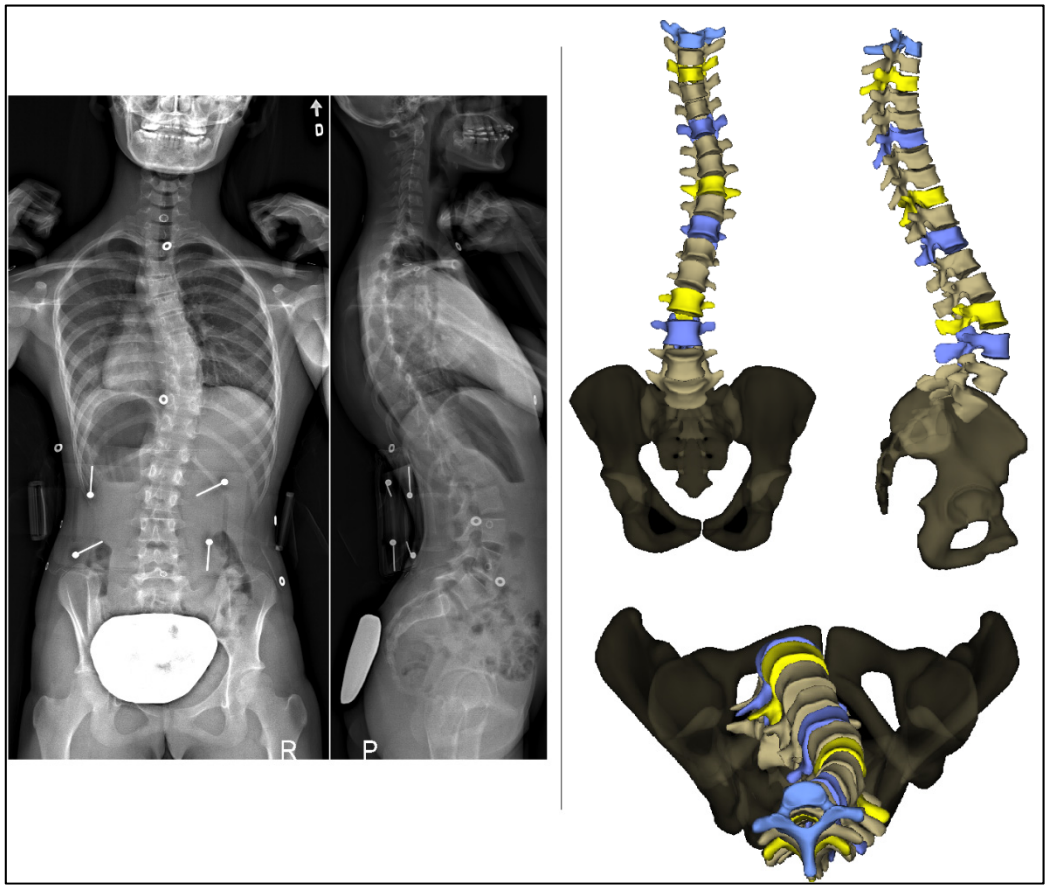


Figure 1.2 Visualisation de la colonne vertébrale par le système EOS de radiographie : dans le plan frontal et sagittal pour les radiographies à gauche, et dans les trois plans de référence pour la reconstruction 3D à droite ; en jaune sont identifiées les vertèbres sommet (apex) et en bleu les vertèbres jonctionnelles, pour le calcul des angles de Cobb

1.2.2 Reconstructions 3D

La colonne vertébrale peut être imagée sous plusieurs modalités, selon l'indication clinique. Le CT et la radiographie à rayons X permettent de visualiser les structures osseuses spécifiquement. Le CT et le système EOS de radiographie permettent d'accéder à des informations en trois dimensions de la colonne, ce qui est très utile pour l'analyse de pathologies comme la scoliose. Le CT est cependant beaucoup plus irradiant que le système EOS (Melhem *et al.*, 2016), et présente l'inconvénient de n'imager les patients qu'en position allongée, qui atténue l'amplitude des courbures de la colonne. Dans ce mémoire nous souhaitons proposer une méthode d'évaluation des techniques de reconstruction 3D de colonne vertébrale à partir du système de radiographie EOS, nous expliquerons donc dans cette partie son fonctionnement.

1.2.2.1 Le système de radiographie EOS

Le système EOS consiste à radiographier simultanément les structures de face et de profil, le patient étant debout dans la cabine de radiographie (voir figure 1.3). Les deux radiographies finales sont ainsi calibrées ensemble, et les structures sont visibles dans la position fonctionnelle du corps.

Figure 1.3 Fonctionnement du système EOS de radiographie : cabine de radiographie (à gauche) et déplacement des tubes à rayons X (à droite)

La méthode de reconstruction 3D consiste à identifier manuellement des repères dans les images, ce qui permet à un modèle statistique de colonne vertébrale de s'initialiser sur les images. Ce modèle 3D statistique est ensuite déformé à l'aide d'outils logiciels afin de faire correspondre au maximum ses projections 2D aux radiographies (Humbert *et al.*, 2009). Une fois ce modèle 3D personnalisé, il est possible d'extraire automatiquement des mesures cliniques permettant l'évaluation des déformations de la colonne vertébrale. La figure 1.4 illustre ce processus de reconstruction 3D pour l'image en vue frontale, mais cette identification se fait bien évidemment sur les deux vues. Le logiciel SterEOS© permet de réaliser ces opérations en 10 à 15 minutes (Jirot *et al.*, 2015; Rehm *et al.*, 2017).

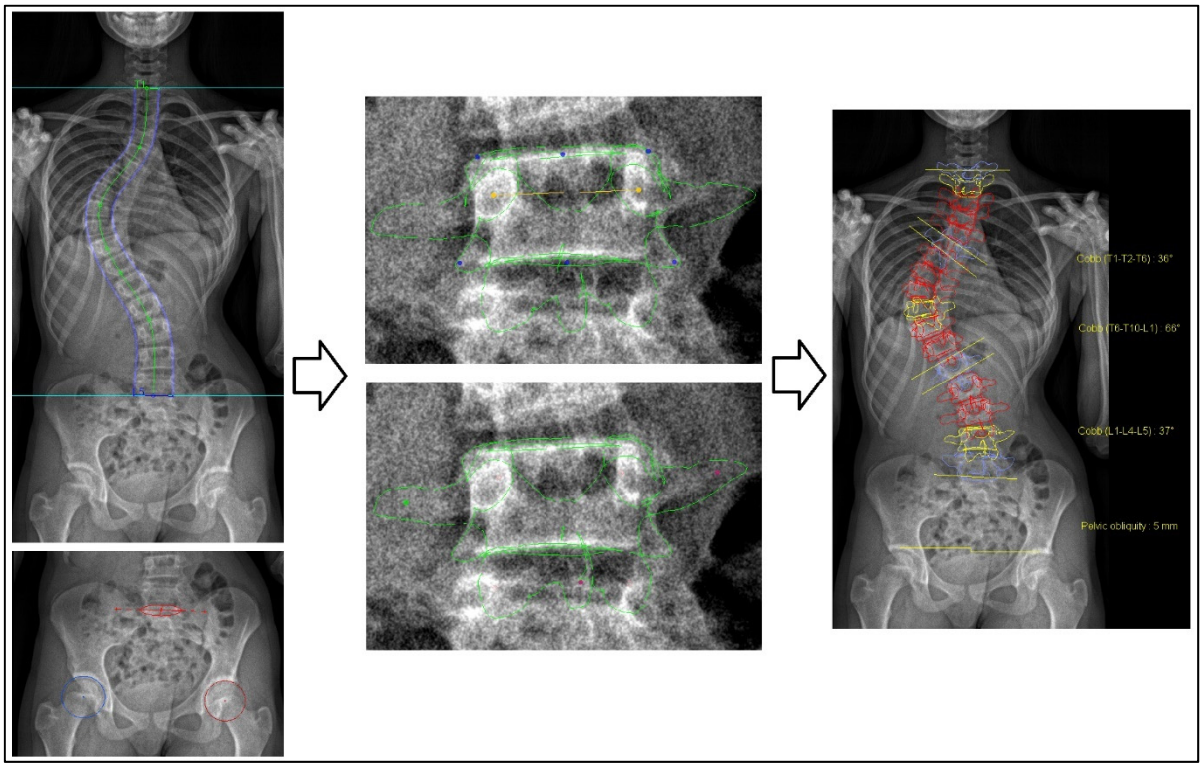


Figure 1.4 Méthode de reconstruction 3D de la colonne vertébrale sous SterEOS© : initialisation du modèle par identification de repères au niveau du pelvis et de l'axe de la colonne (à droite) ; personnalisation du modèle par ajustements des projections 2D vertèbre par vertèbre (au milieu) ; projection 2D finale du modèle 3D avec mesures cliniques

1.2.2.2 Mesures cliniques classiques

Les mesures cliniques pour analyser les déformations rachidiennes les plus utilisées sont les suivantes (Ilharreborde *et al.*, 2011; O'Brien & Spinal Deformity Study Group, 2004) (voir figure 1.5):

- les mesures de cyphoses et de lordoses (T1/T12 ; T4/T12 ; L1/L5 ; L1/S1), qui sont des angles permettant de mesurer la courbure de la colonne dans le plan sagittal ;
- les angles de Cobb, qui sont des angles permettant de mesurer la courbure de la colonne dans le plan frontal (Le plus grand angle de Cobb observé est souvent utilisé comme un indicateur de gravité) ;
- les paramètres pelviens, qui sont des angles au niveau du pelvis permettant de mesurer sa morphologie et sa position par rapport au sacrum (Incidence Pelvienne ; Inclinaison Pelvienne ; Pente Sacrée) ;
- les mesures de rotations vertébrales pour chaque vertèbre dans le plan transverse ;
- l'axe vertical sagittal, qui est une mesure de distance pour évaluer l'inclinaison de la colonne vers l'avant ou vers l'arrière dans le plan sagittal.

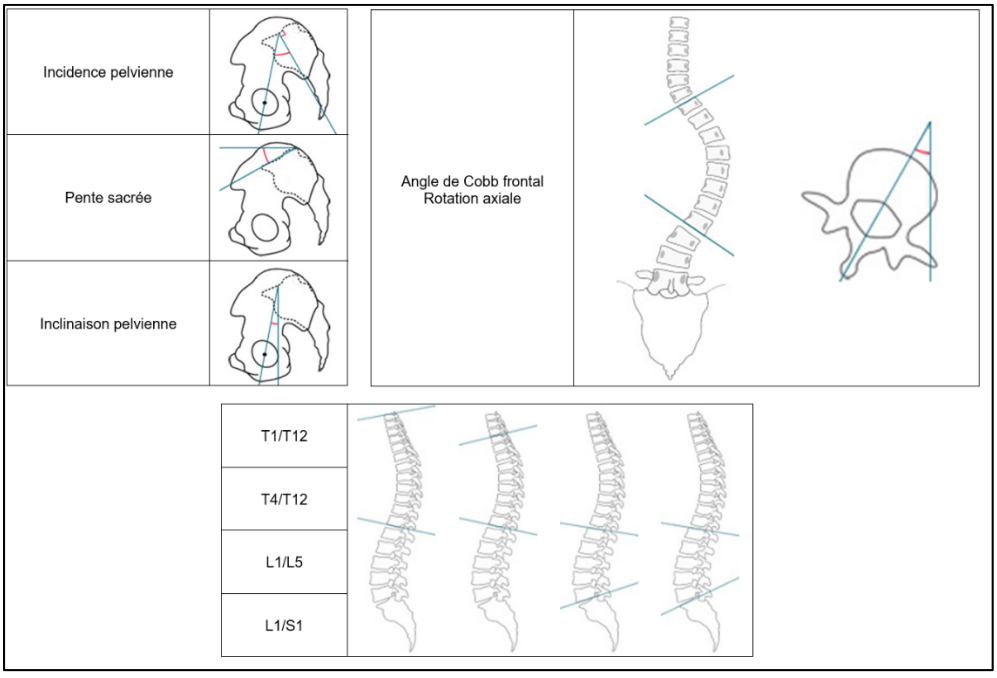


Figure 1.5 Schémas pour le calcul des mesures cliniques sous SterEOS©

1.3 La génération de référence

Une référence sert à évaluer les performances d'une méthode d'analyse d'images, notamment sa précision. L'idéal pour une évaluation serait d'avoir un étalon or. Si ce n'est pas possible, Cardoso *et al.* (2014) décrit la notion de réalité terrain, qui correspond à des analyses ayant fait le plus de consensus ou des analyses avec des valeurs dans lesquelles le plus de confiance a été accordée. La réalité terrain peut être ainsi une référence, même si à l'inverse de l'étalon or, sa précision ne peut pas être complètement vérifiée.

Pour l'évaluation des méthodes de reconstruction 3D de colonne vertébrale, plusieurs types de références ont été utilisées, de l'étalon or à la réalité terrain, que nous allons explorer dans cette partie. Le CT peut être utilisé avec des fantômes (par exemple des bustes artificiels contenant des densités différentes à l'image d'un buste humain), ou avec des sujets humains, à ceci près que le CT est une technologie très irradiante et que la position pour un sujet humain est allongée dans la machine. Les données issues d'analyses humaines, par exemple des mesures manuelles sur images, peuvent être considérées comme réalité terrain si elles font la preuve d'un consensus ou d'une grande confiance parmi des experts (Cardoso *et al.*, 2014). Or il n'existe pas de recommandations de bonnes pratiques sur comment utiliser les performances humaines pour construire une référence. Enfin, il existe d'autres méthodes de production de réalité terrain, faisant appel à des outils de simulations d'images avec des caractéristiques connues, qui permettent de réaliser des évaluations de méthodes d'analyses d'images sans avoir besoin de recruter des sujets humains notamment (Dewalle-Vignion *et al.*, 2015; Stute *et al.*, 2011).

1.3.1 Les fantômes physiques

Un fantôme est un objet aux propriétés connues, permettant de modéliser la réalité. Les fantômes physiques permettent de modéliser la réalité sous la forme d'un objet, par exemple un buste artificiel contenant plusieurs compartiments pour simuler les différentes densités des organes rencontrés dans un buste humain, ou bien dans notre domaine d'application, une colonne vertébrale articulée artificielle (voir figure 1.6) (Chung *et al.*, 2018). Les fantômes physiques permettent d'incorporer les caractéristiques du système d'imagerie, mais ne

permettent pas de reproduire la totalité des caractéristiques que l'on peut rencontrer en pratique (les artéfacts causés par la présence de matériel métallique ou des gaz, la superposition d'autres structures sur la colonne comme les os des bras (voir figure 1.7)). Les déformations pathologiques sont en outre difficiles à simuler avec un fantôme physique, par exemple les cas où les vertèbres sont déformées, de même que les variabilités anatomiques inter-individuelles, avec des sujets qui ont parfois une vertèbre en plus que d'autres. Les fantômes constituent cependant une approche incontournable pour l'évaluation de méthodes d'analyse d'image car leurs caractéristiques sont connues et contrôlées.

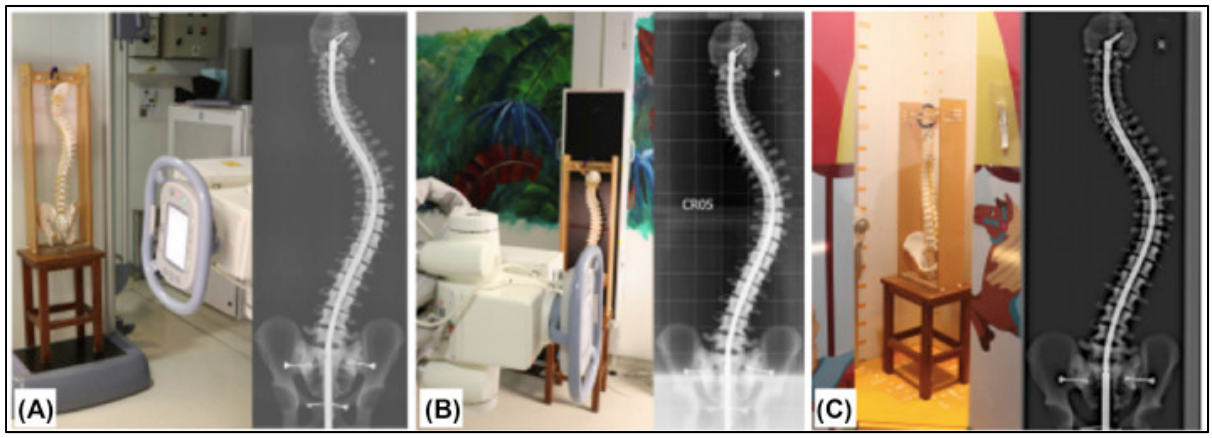


Figure 1.6 Radiographie d'un fantôme physique de colonne vertébrale, en radiologie conventionnelle (A et B), et avec le système EOS de radiographie (C)

1.3.2 Les simulations

Les simulations permettent de faire un compromis entre le manque de représentativité des fantômes physiques et le besoin d'avoir des scènes le plus proches possibles d'une réalité terrain (Dewalle-Vignion *et al.*, 2015). Cependant, les simulations logicielles peuvent encore difficilement prendre en compte tous les paramètres qui peuvent affecter l'image, et ne sont encore que des approximations de ce que l'on peut observer sur le terrain (Despotović *et al.*, 2015).

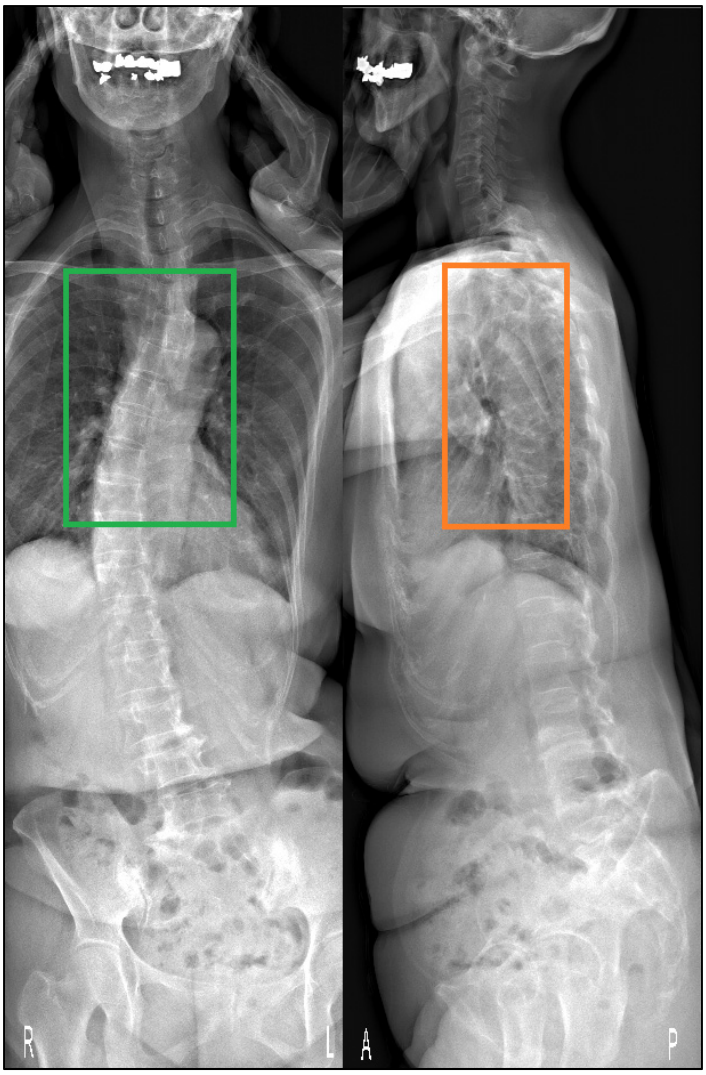


Figure 1.7 Images radiographiques de colonne vertébrales de face et de profil ; dans l'encadré vert : superposition des structures du buste de face (vertèbres, cœur et gros vaisseaux, clavicules, sternum) ; Dans l'encadré orange : superposition des structures du buste de profil (épaule, vertèbres, gros vaisseaux)

1.3.3 L'expertise humaine

Un expert est une personne ayant acquis une grande expérience dans son domaine. En pratique, en analyse d'images médicales, ce sont souvent des médecins, des radiologues ou des chirurgiens qui sont sollicités pour analyser des images et fournir une référence. Pour Heimann *et al.* (2009), même s'il ne s'agit pas d'un étalon or au sens vrai du terme, l'appel à

l'expertise humaine reste le moyen le plus objectif d'obtenir une réalité terrain, et la manière la plus simple d'évaluer les performances d'un algorithme d'analyse d'image (c'est-à-dire de le comparer à une performance humaine). Akhondi-Asl & Warfield (2019) appuient ces propos, en statuant que même si la méthode la plus appropriée pour comparer un résultat d'expert et un résultat d'algorithme n'a pas été mise en évidence, la finalité la plus acceptable serait d'arriver à dire qu'un résultat d'algorithme est suffisamment similaire aux résultats des experts pour considérer l'algorithme comme un substitut acceptable aux experts. Cependant, il n'existe pas de recommandations de pratiques sur la mobilisation d'experts pour la production de référence, si ce n'est que Cardoso *et al.* (2014) décrit la réalité terrain comme un substitut à l'étalon or, du moment qu'un haut niveau de confiance a été démontré, sous la forme d'un consensus d'experts par exemple.

1.3.3.1 L'expert isolé

Maier-Hein *et al.* (2018) a montré que les résultats d'évaluation entre plusieurs méthodes d'analyse d'image pouvaient être impactés très fortement selon l'expert ayant produit la référence. La figure 1.8 résume trois observations de Maier-Hein *et al.* (2018) : le changement de classement des méthodes évaluées selon l'observateur ayant fourni la référence (figure 1.8a et b), selon la métrique utilisée pour évaluer la performance des méthodes évaluées, et selon la méthode utilisée pour combiner les avis de différents observateurs pour construire la référence (figure 1.8c). Les trois observations révèlent un manque de robustesse des classements des méthodes évaluées. Il est en effet impossible avec un seul expert d'identifier s'il y a des erreurs dans la référence qui pourraient être dues à des facteurs humains comme le manque d'expérience sur certains types d'image, la fatigue au moment de l'analyse des images, la maîtrise des outils pour analyser l'image. L'expertise est en outre rarement décrite. Maier-Hein *et al.* (2018) a identifié dans leur méta-analyse de 150 défis d'analyses d'images que 19% des défis n'ont pas précisé le niveau d'expertise des experts mobilisés, seulement 28% des défis ont donné des informations sur les possibles sources d'erreurs dans les références, et seulement 34% des défis ont donné des informations sur la méthode d'analyse des images par les experts humains.

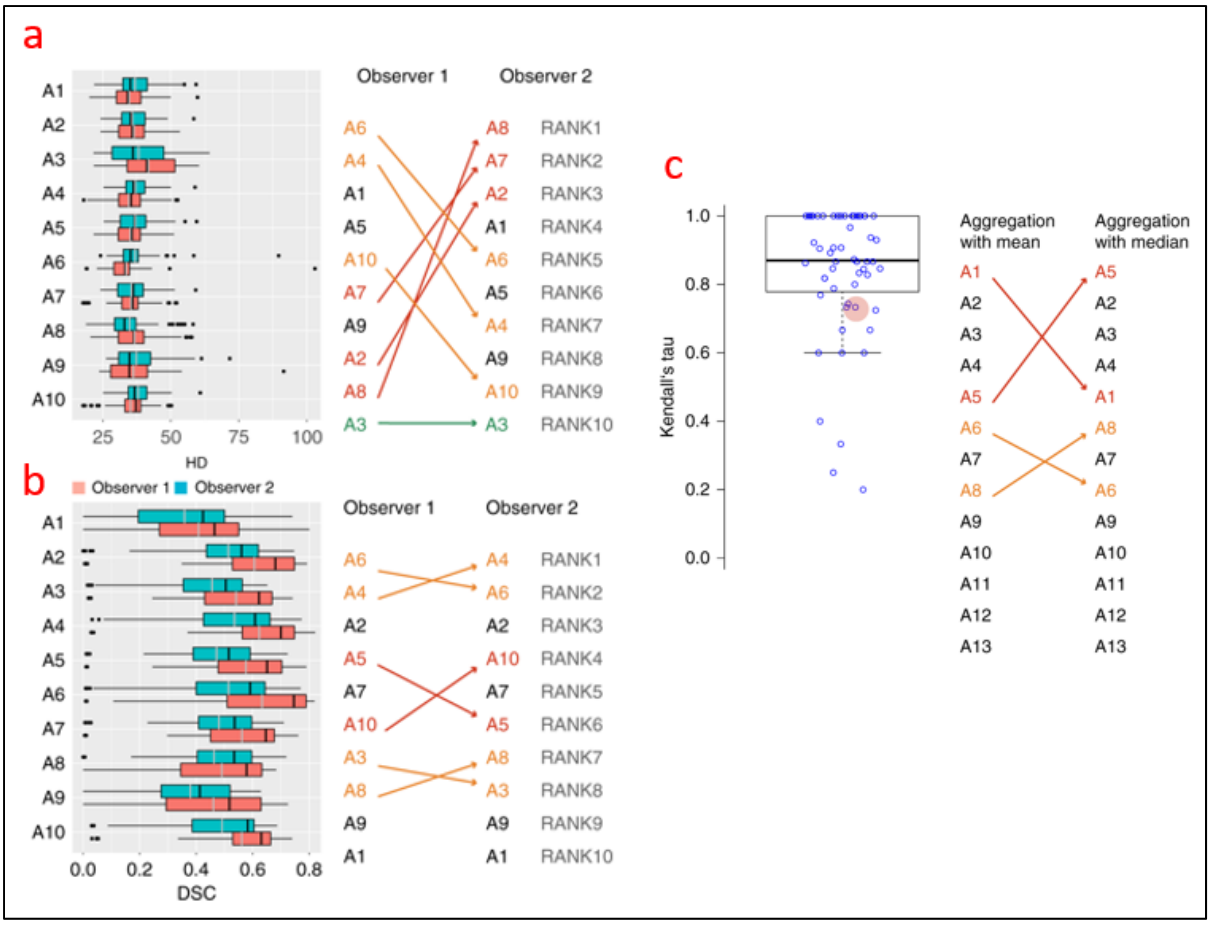


Figure 1.8 (a ; b) Différences de classement de 10 méthodes d'analyse d'image selon deux métriques différentes et selon deux experts différents ayant construit la référence d'évaluation ; (c) Différence de classement de 13 méthodes d'analyse d'images selon deux méthodes d'agrégation de référence experts différentes
Tiré de Maier-Hein *et al.* (2018 p. 5)

1.3.3.2 Les experts ensemble

Lorsque plusieurs experts sont mobilisés pour construire une référence, il est possible d'observer des divergences conséquentes dans les analyses de chacun, ce qui contribue grandement à la difficulté d'identifier la réalité terrain (voir figure 1.9). Il n'existe pas aujourd'hui de recommandations sur comment agréger des opinions d'experts afin de bâtir une référence. Dans l'analyse de Maier-Hein *et al.* (2018), le nombre d'experts mobilisés n'était pas précisé dans 62% des défis de segmentation d'image. Certains résultats

d'évaluations ont vu le classement final changer selon la méthode d'agrégation des annotations expertes (figure 1.8c). Nous exposons cinq méthodes d'agrégation d'opinions experts identifiés dans la littérature : la moyenne, le vote majoritaire, le STAPLE, le consensus et les intervalles de confiance.

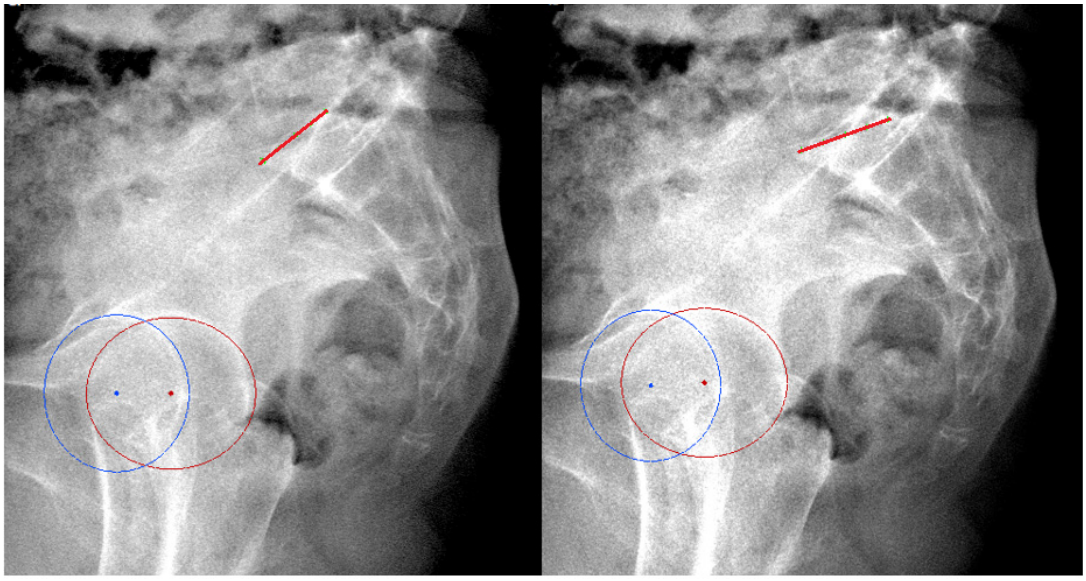


Figure 1.9 Identification de la pente sacrée (ligne rouge) sur des images radiographiques par deux experts différents ; L'expert en (a) a obtenu une pente sacrée de 38 degrés, alors que l'expert en (b) en obtient une à 19 degrés, pour le même patient et la même radiographie

- **Moyenne d'experts**

Udupa *et al.* (2006) propose d'agréger les avis experts en calculant une moyenne des résultats experts. Cela suppose que le comportement de chacun des experts est homogène pour une référence de qualité, mais un seul avis divergent peut grandement modifier la moyenne selon le nombre d'experts mobilisés, or le nombre idéal d'experts pour bâtir une référence est inconnu.

- **Vote majoritaire**

Heimann *et al.* (2009) proposent un système de vote majoritaire. Ils ont créé une référence condensant les résultats de plusieurs algorithmes différents pour lesquels il y avait un

maximum de consensus (si un pixel était considéré "foie" par 5 ou plus des algorithmes évalués, il était considéré "foie" dans le résultat du vote majoritaire, sans justification toutefois de ce nombre de 5). Une solution similaire pour générer une référence "expert" pourrait être une bonne piste, mais Akhondi-Asl & Warfield (2019) soulignent le fait que cette approche n'apporte toujours aucune information quant au nombre d'experts qui doivent être d'accord pour accepter le résultat. De plus, ce type de stratégie ne discrimine pas les experts entre eux, il n'y a ainsi aucune notion de variabilité de performance entre les experts.

- **STAPLE (**Simultaneous Truth And Performance Level Estimation)

Pour tenter de définir la réalité terrain à l'aide d'experts, Warfield *et al.* (2004) introduit le STAPLE qui est un algorithme qui crée une estimation probabiliste de la réalité terrain à partir de plusieurs entrées d'experts. Cet algorithme permet également de fournir des mesures de performance de chacune des entrées pour créer une carte de probabilité. Appliquée à la construction de référence pour l'évaluation de méthodes de segmentation d'images, cette méthode de production de référence consiste à combiner plusieurs segmentations expertes pour générer une réalité terrain où chaque pixel de l'image aura une probabilité d'appartenance à une structure. Warfield *et al.* (2004) explique en outre que STAPLE est capable d'identifier une segmentation correcte même lorsque la majorité des segmentations d'entrée contiennent des erreurs répétées. L'algorithme STAPLE a été présenté par comparaison à des segmentations réalisée par un seul expert, à des fantômes, mais aussi à un groupement d'étudiants en médecine. De ce dernier groupement, une segmentation d'un

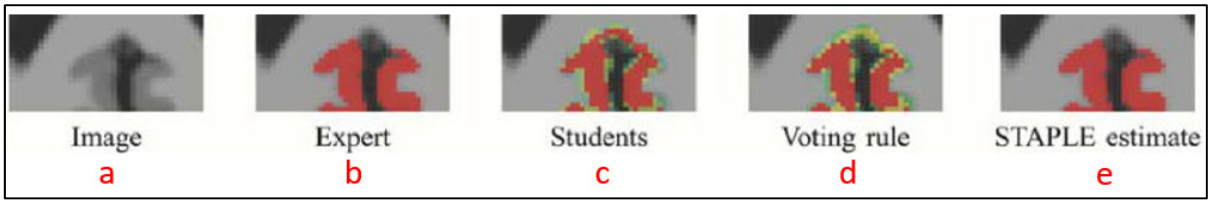


Figure 1.10 Différents résultats de segmentation pour la validation de l'algorithme STAPLE : (a) image native (structure cérébrale), (b) segmentation de l'expert « parfait », (c) segmentations des trois étudiants en médecine, (d) segmentation issue du vote majoritaire, et (e) estimation STAPLE de la réalité terrain
Adapté de Warfield *et al.* (2004, p. 33)

« expert parfait » a été extraite (figure 1.10b). Cette segmentation, nommée dans le document aussi « consensus », a été construite sur la base des trois segmentations des étudiants, après analyses des points de divergences et identification des raisons de désaccords. Le résultat de segmentation parait très similaire à la segmentation proposée par le STAPLE (figure 1.10e), et beaucoup moins à la segmentation des trois étudiants isolés ou du vote majoritaire (figure 1.10c et 1-10d). Le STAPLE présente des limitations, parmi lesquelles la difficulté de l'adapter à d'autres domaines d'analyse d'images, le grand nombre de paramètres à ajuster (Dewalle-Vignion *et al.*, 2015), ainsi que la faiblesse des preuves apportées sur son intérêt par rapport à un consensus d'expert.

- **Consensus**

Le concept de l'expert « parfait » avancé par Warfield *et al.* (2004) est un consensus d'experts, puisque les désaccords sont discutés afin de trouver un compromis et délivrer un résultat de segmentation unique. Ils n'ont toutefois pas investigué cette méthode pour produire des références, seulement pour réaliser des observations quant à la performance du STAPLE pour produire des segmentations. Cela rejoint l'expérience de terrain de EOS Imaging, et de leur processus de reconstruction 3D de la colonne vertébrale. Ce processus est résumé dans la figure 1.11 : un premier expert procède à la reconstruction 3D depuis la paire d'images, puis un deuxième expert revoit et corrige si besoin la reconstruction, et en cas de

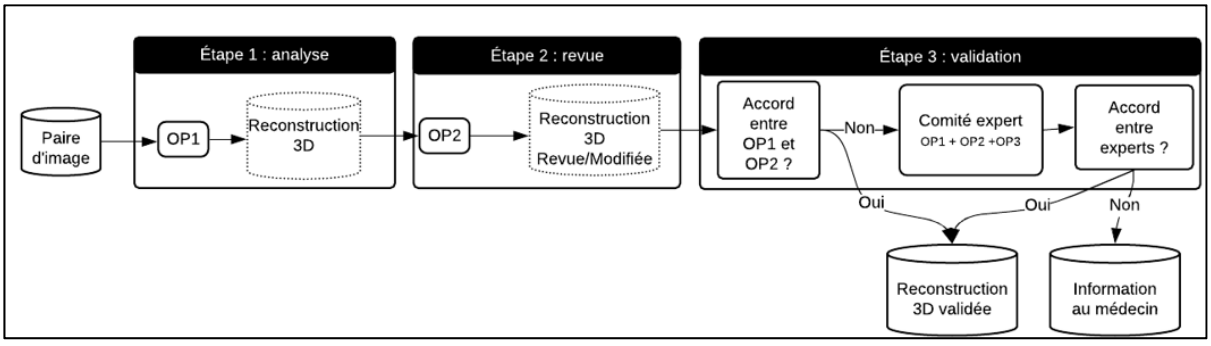


Figure 1.11 Processus de reconstruction 3D mis en place par l'entreprise partenaire EOS Imaging, incluant des discussions entre experts pour valider la reconstruction 3D

modifications un accord est recherché entre ces deux experts. Si l'accord n'est pas trouvé, il est fait appel à un troisième expert pour intégrer la discussion sur les points de désaccords, et si un compromis n'est toujours pas trouvé sur la reconstruction 3D, une note est adressée au médecin prescripteur pour l'informer des points de désaccords et des incertitudes observées.

- **Intervalles de confiance**

Aubert *et al.* (2019) utilise un autre moyen pour agréger les opinions de plusieurs experts, en utilisant des intervalles de confiances de mesures cliniques issues des reconstructions 3D de colonne vertébrale pour pouvoir évaluer la performance d'une méthode automatique de reconstruction 3D. Avec 3 experts indépendants ayant procédé chacun une fois à la reconstruction 3D de chaque paire d'image, ils proposent d'analyser les résultats de sa méthode automatique selon la position de ses mesures cliniques par rapport aux intervalles de confiance expert (voir figure 1.12), calculés avec les écarts types observés sur les mesures.

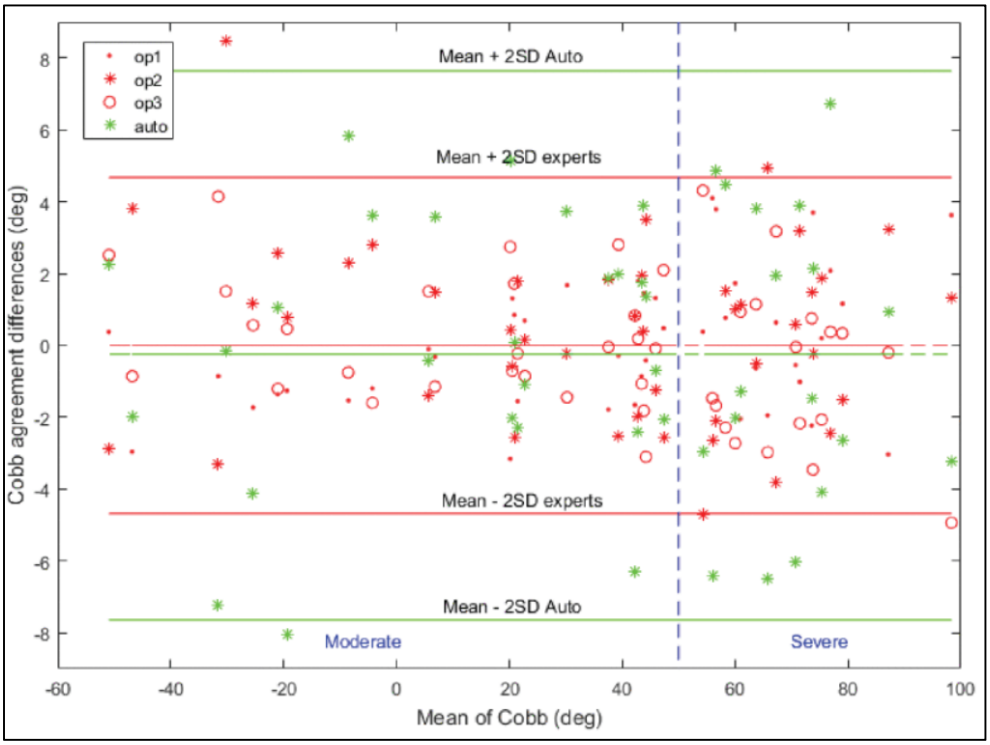


Figure 1.12 Graphique de Bland-Altman montrant la distribution des mesures cliniques de l'angle de Cobb entre un groupement d'experts (en rouge) et un algorithme automatique (en vert), pour la reconstruction 3D de colonne vertébrale, et les intervalles de confiance pour le groupe d'expert et pour l'algorithme automatique (moyenne ± 2 écarts types).
Tiré de Aubert *et al.* (2019)

Cette méthode rejoint des travaux cliniques comme ceux de Ilharreborde *et al.* (2011), qui se sont appuyés sur des calculs de reproductibilités pour comparer la fiabilité des mesures cliniques avant et après chirurgie.

1.3.4 Conclusion

Construire une référence en l'absence d'étalon or est une tâche complexe pour laquelle plusieurs solutions ont été proposées dans la littérature. Pour la reconstruction 3D de la colonne vertébrale, le CT est difficilement envisageable, au regard de l'exposition radiologique et de la position allongée. L'utilisation de fantômes présente la limitation de la représentativité des images, que ce soit sur les pathologies en soit (les déformations du fantôme sont limitées) ou sur la qualité image (présence de gaz, de superpositions de structures). Un expert isolé, une moyenne d'experts ou un vote majoritaire, sont des moyens de production de référence qui présentent des biais humains difficilement quantifiables, comme la fatigue, le manque d'expérience, ou les erreurs humaines. Les intervalles de confiances pour des mesures dans les images présentent des biais similaires, avec la possible présence de valeurs aberrantes dues aux biais humains. Enfin, nous avons vu que la méthode du STAPLE semble permettre d'obtenir le même résultat qu'un consensus d'experts, sans que des analyses aient été réalisées par Warfield *et al.* (2004) sur le nombre d'experts à mobiliser et sur l'impact de la qualité de l'image pour la création de l'expert « parfait ».

Comme décrit par Cardoso *et al.* (2014), une réalité terrain doit pourtant être une référence ayant le plus haut niveau de confiance, à l'image d'un consensus. Il serait intéressant dans ce sens de savoir si le consensus d'experts permet de diminuer les écarts d'opinions entre experts et les biais humains, et d'en apporter la preuve pour la reconstruction 3D de colonne vertébrale. Nous proposons dans ce mémoire d'associer le concept du consensus, où des experts discutent de leurs points de désaccords, à l'utilisation des intervalles de confiance de Aubert *et al.* (2019) afin de construire une référence de reconstruction 3D de colonne vertébrale.

Pour évaluer des méthodes d'analyse d'images, obtenir une référence ne suffit pas, il faut inclure à l'évaluation des métriques afin de quantifier les erreurs entre la méthode évaluée et la référence.

1.4 Quantification des écarts à la référence

Pour quantifier la performance d'un algorithme d'analyse d'images, il est nécessaire d'utiliser des mesures objectives qui soient capables de représenter ou de quantifier les erreurs. Taha & Hanbury (2015) ont regroupé une vingtaine de métriques, les plus couramment utilisées dans la littérature pour des évaluations de méthodes de segmentation 3D d'images médicales, en essayant de donner une définition standard de chacune. Ils proposent en outre une catégorisation de ces métriques selon leurs natures et leurs définitions : les métriques de recoupement dans l'espace, les métriques basées sur les volumes, et les métriques basées sur les distances dans l'espace. Nous ajoutons à cela des métriques permettant de comparer des mesures sur des images, comme des paramètres cliniques en radiologie par exemple. En effet, nous souhaitons mettre en place une méthode d'évaluation des reconstructions 3D de colonne vertébrale, qui permette à la fois d'évaluer la précision des modèles 3D et des indices cliniques issus de ces modèles.

Pour définir ces métriques, Taha & Hanbury (2015) proposent des calculs en fonction de la matrice de confusion (voir figure 1.13) :

- les vrais positifs (TP) : ensemble de pixels inclus dans la segmentation issue de la méthode évaluée et appartenant vraiment à l'objet dans la réalité terrain ;
- les vrais négatifs (TN) : ensemble de pixels exclus de la segmentation issue de la méthode évaluée et n'appartenant pas non plus à l'objet dans la réalité terrain ;
- les faux positifs (FP) : ensemble de pixels inclus dans la segmentation issue de la méthode évaluée et n'appartenant pas à l'objet dans la réalité terrain ;
- les faux négatifs (FN) : ensemble de pixels exclus de la segmentation issue de la méthode évaluée et appartenant pourtant à l'objet dans la réalité terrain.

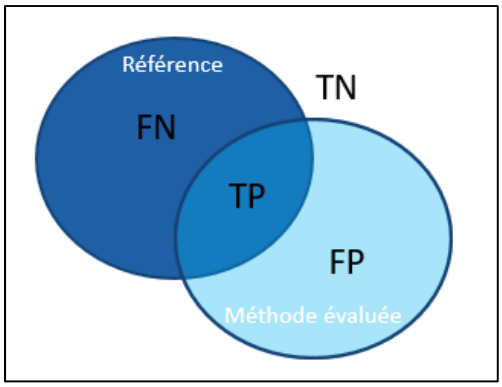


Figure 1.13 Illustration de la matrice de confusion

1.4.1 Les recoupements dans l'espace :

- **Indice de Dice**

Le coefficient de Dice (équation 1.1), aussi appelé indice de recoupement, est une des métriques les plus utilisées pour la validation de segmentations volumiques d'images médicales (Taha & Hanbury, 2015). Il permet de comparer directement deux segmentations et d'évaluer la reproductibilité des segmentations. Il est accepté qu'un Dice supérieur à 80% reflète un accord presque parfait entre deux segmentations. Cet indice de similarité peut cependant surestimer la vraie valeur de similarité, il est donc utile pour comparer des résultats mais pas pour les valider (Cárdenes *et al.*, 2009).

$$DICE = \frac{2TP}{2TP + FP + FN} \tag{1.1}$$

- **Indice de Jaccard**

L'indice de Jaccard (JI) est également très utilisé. Il permet de calculer la surface en commun de deux objets par rapport à leurs surfaces totales (équation 1.2). Cet indice n'apporte cependant pas plus d'information que le coefficient de Dice, à cause de la relation non linéaire entre les deux métriques (Akhondi-Asl & Warfield, 2019; Taha & Hanbury, 2015). L'erreur de recoupement volumique (VOE) est issue de cet indice, et adapté pour des mesures volumétriques (Heimann *et al.*, 2009; Laurent *et al.*, 2016).

$$JI = \frac{TP}{TP + FP + FN} \tag{1.2}$$

- **Sensibilité et Spécificité**

La sensibilité (Se), est le pourcentage de vrais positifs, et donc la probabilité d'inclure un pixel à la segmentation sachant qu'il appartient réellement à l'objet (équation 1.3). La spécificité (Sp), est le pourcentage de faux négatifs, c'est-à-dire la probabilité de ne pas inclure un pixel à la segmentation sachant qu'il n'appartient pas à l'objet (équation 1.4). D'après (Taha & Hanbury, 2015), ces deux mesures ne sont pas communes dans la littérature, à cause de leur sensibilité à la taille des segmentations : plus la segmentation est de grande taille, moins les erreurs seront discriminées.

$$Se = \frac{TP}{TP + FN} \tag{1.3}$$

$$Sp = \frac{TN}{TN + FP} \tag{1.4}$$

- **Courbes ROC et aire sous la courbe**

Une manière de caractériser les performances d'un test est de construire la courbe ROC (Receiver Operating Characteristic), qui exprime le taux de vrais positifs en fonction du taux de faux positifs. Taha & Hanbury (2015) donnent une définition de l'AUC (aire sous la courbe) en fonction des cardinalités de précision (équation 1.5)

$$AUC = 1 - \frac{1}{2}\left(\frac{FP}{FP + FN} + \frac{FN}{FN + TP}\right) \tag{1.5}$$

1.4.2 Les volumes :

- **Similarité volumique**

Taha & Hanbury (2015) définissent la similarité volumique (VS), qui est la différence de volume absolue (équation 1.6). Le recoupement des volumes n'est cependant pas pris en compte dans cette métrique.

$$VS = \frac{|FN - FP|}{2TP + FP + FN} \tag{1.6}$$

- **La différence de volume relative**

La différence de volume relative (RVD) est utile pour comparer un volume de segmentation $V_{méthode\ évaluée}$ à un volume de référence $V_{référence}$ (équation 1.7). Cette formule est également applicable aux surfaces fermées avec la différence de surface relative (RSD) (Heimann *et al.*, 2009).

$$RVD = \frac{\left|V_{méthode\ évaluée} - V_{référence}\right|}{V_{référence}} \times 100 \tag{1.7}$$

1.4.3 Les distances dans l'espace :

Ces métriques sont les plus utilisées en évaluation de segmentation d'image, et se basent sur la position des objets dans l'espace. Lorsque les contours des segmentations sont les aspects les plus importants à étudier, ces métriques sont à privilégier.

- **Distance symétrique moyenne**

La distance symétrique moyenne (ASD) permet de calculer la distance moyenne entre la surface 3D segmentée à évaluer A et la surface 3D segmentée de référence B (Heimann *et al.*, 2009) (équation 1.8).

$$ASD(A,B) = \frac{1}{|S(A)| + |S(B)|}\left(\sum_{s_A \in S(A)} d\big(s_A, S(B)\big) + \sum_{s_B \in S(B)} d\big(s_B, S(A)\big)\right) \tag{1.8}$$

Avec S(A) l'ensemble des voxels de surface de A

Et $d\big(s, S(A)\big) = min_{s_A \in S(A)} \|s - s_A\|$ la distance la plus courte à une surface s.

- **Distance de Hausdorff**

La distance de Hausdorff (HD) (Huttenlocher *et al.*, 1993) est la valeur symétrique maximale que peut prendre la distance entre la surface 3D de segmentation A et la surface 3D de référence B. Associée à l'ASD, la HD permet d'avoir des informations sensibles sur les valeurs aberrantes, et de les quantifier (équation 1.9). Pour Heimann *et al.* (2009), cette métrique est d'extrême importance dans les planifications chirurgicales où identifier la plus grande erreur est plus important qu'identifier une moyenne des erreurs.

$$HD(A,B) = \max\left\{max_{S_A \in S(A)} d(s_A, S(B)), max_{S_B \in S(B)} d(s_B, S(A))\right\} \quad (1.9)$$

Avec S(A) l'ensemble des voxels de surface de A

- **Distance moyenne quadratique**

Dérivée de l'ASD, la distance moyenne quadratique (RMSD) est aussi basée sur les distances entre deux surfaces 3D A et B (équation 1.10). Pour évaluer la précision des segmentations la RMSD est plus pertinente que l'ASD car les erreurs sont amplifiées (Heimann *et al.*, 2009).

$$RMSD(A,B) = \sqrt{\frac{1}{|S(A)| + |S(B)|}} \times \sqrt{\sum_{s_A \in S(A)} d^2\left(s_A, S(B)\right) + \sum_{s_B \in S(B)} d^2\left(s_B, S(A)\right)} \quad (1.10)$$

1.4.4 Les mesures dans l'image

- **Taux de succès**

Sur un échantillon de mesures sur un nombre n d'images, le taux de succès (équation 1.11) correspond au nombre de mesures situées dans l'intervalle de confiance de référence (Aubert *et al.*, 2019).

$$Taux\ de\ succès = 100 \times \frac{1}{n} \sum_{n} Succès \quad (1.11)$$

Avec Succès = 1 si $Mesure_n \in$ ICmesure, sinon Succès = 0

- **MAD et SDAD**

La moyenne des différences absolues (MAD) (Galbusera *et al.*, 2019; Korez *et al.*, 2020) permet de mesurer la moyenne des écarts absolus entre la valeur obtenue par la méthode évaluée et la valeur de la référence pour un échantillon de mesures sur un nombre n d'images (équation 1.12). La déviation standard des différences absolues (SDAD) peut également être calculée (équation 1.13).

$$MAD = \frac{1}{n}\sum_{n}|Valeur\ de\ la\ référence - \ Valeur\ de\ la\ méthode\ évaluée| \tag{1.12}$$

$$SDAD = \sqrt{\frac{1}{n}\sum_{n}(|Valeur\ de\ la\ référence - \ Valeur\ de\ la\ méthode\ évaluée|)^2} \tag{1.13}$$

- **SMAPE**

L'erreur de pourcentage absolu moyen symétrique est une mesure d'exactitude (Kim *et al.*, 2020; Wang *et al.*, 2019). Elle peut être calculée pour une seule mesure ou pour plusieurs mesures combinées sur un nombre n d'images (équation 1.14).

$$SMAPE = \frac{1}{nombre\ de\ mesures}\sum_{n}^{nombre\ de\ mesures}\frac{|Valeur\ de\ la\ référence_n - \ Valeur\ de\ la\ méthode\ évaluée_n|}{|Valeur\ de\ la\ référence_n + \ Valeur\ de\ la\ méthode\ évaluée_n|} \tag{1.14}$$

1.4.5 Conclusion

Pour l'évaluation des méthodes de reconstruction 3D de colonne vertébrale, plusieurs métriques semblent adaptées, parmi lesquelles les métriques de volume, de distance et les métriques de performance de mesures, puisque la reconstruction 3D inclut à la fois des objets 3D surfaciques et des mesures cliniques (voir figure 1.14). La précision des mesures cliniques et des modèles 3D, en dépit du manque d'étalon or pour la position debout, a été analysée avec l'apport de preuves de précision et de fiabilité par comparaison à des CT de fantômes physiques ou de régions restreintes du rachis (Melhem *et al.*, 2016). Nous allons

nous concentrer dans un premier temps sur les mesures cliniques pour la mise en place de notre méthode d'évaluation. Dans cette optique, le MAD et le SMAPE permettent d'évaluer les proportions d'erreurs dans un lot de données de mesures dans des images, nous pourrons ainsi les utiliser pour nos évaluations de performances. Il nous faudra par la suite conduire une analyse sur le lien entre les mesures cliniques et les modèles 3D afin de compléter notre méthode d'évaluation. En effet, si le niveau de confiance atteint par nos mesures cliniques de référence est le même pour les points des modèles 3D, nous serons capables de les utiliser comme référence et donc d'utiliser davantage de métriques d'évaluation.

Il nous reste dans cette revue à analyser maintenant quels sont les critères d'évaluations les plus utilisés dans la littérature, afin de pouvoir donner un sens aux erreurs d'analyses.

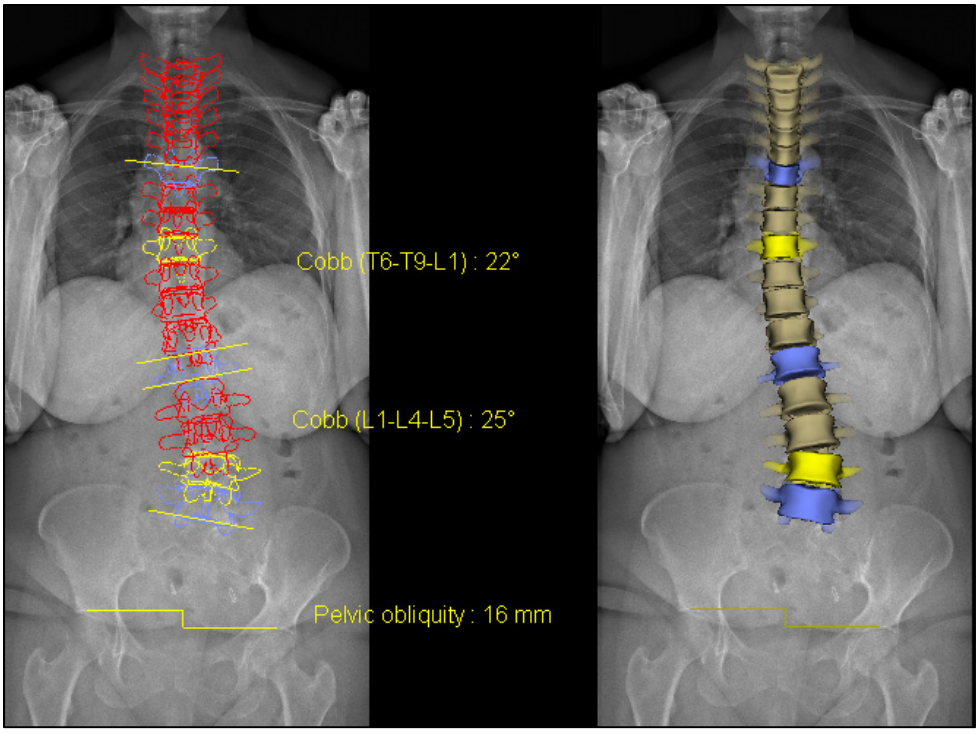


Figure 1.14 Reconstruction 3D de colonne vertébrale : mesures cliniques (à gauche) et modèle 3D (à droite)

1.5 Qualification des écarts à la référence

1.5.1 Qualifier la performance

Dans la littérature trois critères principaux se dégagent pour évaluer la performance d'une méthode d'analyse d'image, à savoir la reproductibilité, la précision et la justesse (figure 1.15). Pour des raisons sémantiques, nous nous rattacherons à la norme ISO 5725-1 pour la définition de ces critères (*ISO 5725-1:1994*, s. d.).

- **La reproductibilité** est définie par « l'étroitesse de l'accord entre les résultats ». Une méthode reproductible a ainsi des erreurs aléatoires qui sont de faible amplitude.
- **La justesse** est « l'accord entre la moyenne arithmétique d'un grand nombre de résultats et la valeur de référence vraie ou acceptée ». La justesse est également exprimée en termes de « biais » ou « d'erreur systématique ».
- **L'exactitude ou la précision** « est utilisée à la fois en référence à la justesse et à la fidélité ». L'exactitude cumule les erreurs systématique et aléatoire en mesurant « le déplacement total d'un résultat par rapport à la valeur de référence ».

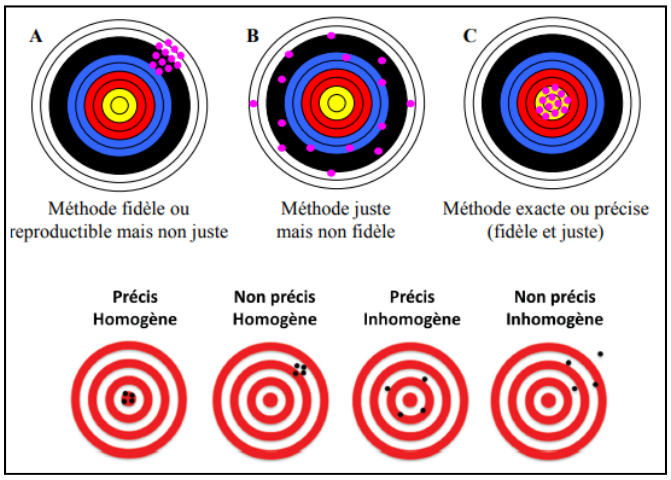


Figure 1.15 Illustration des concepts de fidélité (homogénéité), justesse et exactitude
Adapté de Humbert *et al.* (2009) et Laurent *et al.* (2016)

Pour la reconstruction 3D de colonne vertébrale, ce sont surtout les concepts de précision et de reproductibilité qui reviennent dans la littérature. Humbert *et al.* (2009) et Nérot *et al.* (2015) ont utilisé la valeur moyenne des distances points-surfaces entre deux formes pour déterminer la précision. D'autres ont estimé la reproductibilité en calculant le RMSSD (équation 1.15), en construisant des intervalles de confiance à 95% pour les mesures cliniques (Humbert *et al.*, 2009; Ilharreborde *et al.*, 2011; Somoskeöy *et al.*, 2012; Nérot *et al.*, 2015).

$$RMSSD = \sqrt{\frac{1}{n}\sum_{n} {SD_n}^2} \tag{1.15}$$

Avec SD^2 la variance des mesures ou carré de l'écart type et RMSSD la moyenne quadratique des écarts types.

1.5.2 Qualifier le comportement

D'autres critères sont évoqués dans la littérature qui paraissent tout aussi pertinents, notamment pour caractériser le comportement de certains algorithmes d'analyse d'image.

- **L'efficience :** ce critère permet de savoir si la méthode évaluée est viable en pratique (Udupa et al., 2006), en prenant en compte par exemple le temps d'analyse par rapport à l'humain.
- **La robustesse :** ce critère permet d'évaluer la capacité d'une méthode à fournir les mêmes résultats malgré des changements de conditions expérimentales (Jannin *et al.*, 2002). Laurent *et al.* (2016) a couplé l'indice de Jaccard à la différence relative de surface (équation 1.16) pour estimer la robustesse.

$$Robustesse = JI \times (100 - |RSD|) \tag{1.16}$$

- **La sensibilité à la sur ou sous segmentation (Θ) :** Laurent *et al.* (2016) a défini la sensibilité comme étant la « tendance globale de l'algorithme à sur ou sous segmenter la forme d'intérêt ». Pour sa définition, un paramètre de sensibilité α est fixé à « 1.96 fois

l'écart type de l'aire de la référence, soit l'intervalle où se situeraient 95% des segmentations expertes, en supposant qu'elles soient normalement distribuées » (équation 1.17). Couplé à la robustesse, Laurent *et al.* (2016) avance qu'il est possible de discriminer deux comportements d'erreurs différents.

$$\boldsymbol{\Theta}\,(\boldsymbol{\alpha}) = 100 \times \frac{Card(\{S_{méthode\ évaluée} : \big||Méthode\ évaluée| - |Référence|\big| < \alpha\})}{Card(\{S_{référence}\})} \qquad (1.17)$$

Avec, $S_{méthode\ évaluée}$ l'ensemble de pixels ou de voxels identifiés par la segmentation à évaluer

$Card(\{S_{méthode\ évaluée} : \big||Méthode\ évaluée| - |Référence|\big| < \alpha\})$ le nombre de pixels ou de voxels dont la position ne dépasse pas le seuil α et

$\alpha = 1.96 \times \mathrm{SD}\ (|Référence\ |)$

- **La sensibilité aux valeurs aberrantes Δ** : Laurent *et al.* (2016) a défini ce critère afin de pouvoir analyser l'impact des valeurs aberrantes sur le comportement des algorithmes. Après le choix d'un seuil de valeur aberrantes à partir de la RMSD, il calcule « *le poids de la variance de la distance entre les valeurs aberrantes par rapport à la variance des distances totales »* (équation 1.18), ainsi le nombre de valeurs aberrantes dans le résultat de la méthode à évaluer a plus d'impact que seulement la position de ces erreurs.

$$\boldsymbol{\Delta} = 100 \times (1 - \frac{Variance\ des\ distances\ aberrantes}{Variance\ des\ distances\ totales}) \qquad (1.18)$$

1.5.3 Conclusion

Les critères d'évaluation permettent de caractériser les types d'erreurs d'analyse ainsi que les comportements des méthodes évaluées pour mieux guider les développements et les améliorations. Nous n'avons pas identifié dans la littérature de méthodes d'évaluations qui inclue les besoins cliniques en analyse d'images médicales, alors que ce type de critère pourrait apporter une information capitale sur la pertinence des méthodes et la pertinence de la recherche constante d'amélioration. En d'autres termes, à quoi bon atteindre une précision

de 100% si cela n'a aucun ou très peu d'impact sur la prise de décision thérapeutique? De façon plus pratique, une méthode précise à 95% peut-elle être considérée pour une application clinique si les 5% d'erreurs concernent les situations pour lesquelles les conséquences des erreurs sont les plus graves?

Pour notre application des reconstructions 3D de la colonne vertébrale, les trois critères principaux que sont la précision, la reproductibilité et la justesse sont applicables, notamment pour les mesures cliniques issues des reconstructions 3D. Dépendamment de comment nous aurons défini nos références d'évaluation, il faudra donc réfléchir à comment adapter ces critères aux données d'évaluations. Certains des critères évoqués sont en effet facilement adaptables pour notre champ d'application, à l'image de la sensibilité à la sur ou sous segmentation.

S'agissant d'une tâche d'analyse médicale, il faudra également conduire une réflexion sur la pertinence des critères retenus, en incluant des aspects cliniques à nos critères de performances.

1.6 Résumé

La revue de littérature nous permet de constater que l'évaluation de méthodes d'analyses d'images est régulièrement sources de travaux, que ce soit au niveau de la construction de référence, de la quantification des écarts à cette référence ou de la qualification de ces écarts.

La recherche de référence en l'absence d'étalon or est encore aujourd'hui une tâche complexe, avec le recours quasi systématique à des experts humains sans qu'il n'y ait de recommandations claires sur la manière de combiner les avis d'experts différents (Maier-Hein *et al.*, 2018). Pour ce faire, nous allons conduire une analyse sur la pertinence du consensus entre experts pour bâtir une référence, en étudiant l'évolution de la confiance dans les mesures cliniques selon l'implémentation ou non de discussions entre différents experts pour la reconstruction 3D de colonne vertébrale.

Le choix des métriques d'évaluations pour quantifier les erreurs d'analyse d'images a été le sujet de recommandations dans la littérature (Taha & Hanbury, 2015), et nous pouvons facilement identifier les plus pertinentes pour notre méthode d'évaluation. Nous faisons le choix de privilégier dans un premier temps les métriques permettant de mesurer la performance de mesures cliniques, et notre méthode pourra évoluer avec l'utilisation de métriques complémentaires si nous apportons la preuve que la référence proposée pour les mesures cliniques est extrapolable aux modèles 3D.

Concernant les critères pour nous aider à qualifier les types d'erreurs, nous souhaitons en plus des critères standards de précision et de reproductibilité, prendre en compte les besoins cliniques de prise en charge des déformations rachidiennes qui sont au cœur de la reconstruction 3D de colonne vertébrale. Il nous apparait primordial de pouvoir construire ou adapter des critères cliniques pour évaluer la performance d'une méthode, afin de pouvoir hiérarchiser les erreurs par gravité des impacts potentiels qu'elles pourraient avoir sur un diagnostic ou sur un choix thérapeutique. Cela permettrait de mieux guider les développeurs de méthodes de reconstruction 3D en leur donnant une information contextuelle qui aiderait à prioriser les perspectives d'améliorations.

Ces réflexions vont dans le sens des recommandations de Maier-Hein *et al.* (2018). Comme vu précédemment, cette revue de 150 défis d'analyses d'images médicales de plusieurs conférences scientifiques internationales a montré ce manque de standardisation dans les différents cadres d'évaluations utilisés, ce qui a engendré des résultats de compétition peu robustes (changements de classements selon les experts ayant réalisé la référence et selon les métriques d'évaluation utilisées). En d'autres termes, nous avons vu dans la littérature qu'il existe de nombreux outils d'évaluations, sur la manière de construire une référence, sur la manière de quantifier et de qualifier les erreurs d'analyses, mais peu de consensus sur ce qu'il faut utiliser en regard de quel besoin.

Dans le travail présenté dans ce mémoire, nous proposons un cadre d'évaluation pour les reconstructions 3D de colonne vertébrale, avec plus particulièrement un travail sur la création

d’une référence à haut niveau de confiance à l’aide d’experts ainsi que sur l’inclusion des aspects cliniques.

CHAPITRE 2

OBJECTIFS ET CONTRIBUTIONS

La revue de littérature nous a permis d'analyser les différentes méthodes d'évaluations en analyse d'images, que ce soit au niveau de la construction de référence, du choix de métriques d'évaluation ou du choix des critères d'évaluation. Dans le cadre plus spécifique de la reconstruction 3D de colonne vertébrale à partir d'images EOS pour des sujets debout, il n'existe pas de recommandations de bonne pratique pour l'évaluation des reconstructions 3D.

Notre objectif principal de recherche est de proposer un cadre d'évaluation pour les méthodes de reconstructions 3D de la colonne vertébrale, incluant un travail sur la recherche d'une méthode de construction de référence et l'apport d'informations cliniques dans la méthode d'évaluation afin de mieux guider les développeurs de nouvelles méthodes de reconstructions 3D. Nos sous objectifs sont donc les suivants :

- **Objectif 1 :** Proposer une nouvelle méthode de création de référence de mesures cliniques fondée sur le processus existant utilisé par l'entreprise partenaire EOS Imaging afin de bâtir une réalité terrain démontrant le plus haut niveau de confiance. La littérature nous a montré les failles des méthodes existantes, notamment sur l'appel à des experts pour construire la référence. Si Warfield *et al.* (2004) proposent un consensus d'expert pour proposer « l'expert parfait » comme référence pour la validation du STAPLE, ils n'en font aucune analyse sur sa performance, ses faiblesses et ses forces. Cardoso *et al.* (2014) ont pourtant décrit la réalité terrain comme un substitut acceptable à l'étalon or, pourvu que le niveau de confiance ait été démontré ou qu'elle soit le résultat d'un consensus. Puisque Maier-Hein *et al.* (2018) ont pointé le manque de recommandations de la littérature sur la construction de référence, et surtout une grande hétérogénéité de pratiques, nous proposons au chapitre 3 une nouvelle méthode de création de référence.

- **Objectif 2 :** Proposer une méthode d'évaluation incluant des informations cliniques sur la performance des méthodes de reconstructions 3D évaluées, permettant de hiérarchiser les erreurs selon leur gravité, et ainsi mieux guider les développements futurs.
- **Objectif 3 :** Analyser l'intérêt de la méthode de construction de référence pour les modèles 3D de colonne vertébrale. Dans le contexte des reconstructions 3D de colonne vertébrale, cela permettra de proposer une recommandation générale sur la construction de référence, afin d'homogénéiser les pratiques de recherche, qui actuellement font appel à des experts isolés et non à un consensus d'expert pour bâtir la référence (Aubert *et al.*, 2019; Kim *et al.*, 2020; Yeung *et al.*, 2020)

CHAPITRE 3

LE CONSENSUS POUR LA PRODUCTION DE RÉFÉRENCE DE RECONSTRUCTIONS 3D DE LA COLONNE VERTÉBRALE

3.1 Introduction

Identifier une référence de qualité en l'absence d'étalon or en imagerie médicale est une tâche à laquelle peu d'études se sont intéressées. Maier-Hein *et al.* (2018) a identifié plusieurs pratiques communes dans la littérature lorsqu'il s'agit de faire appel à des experts pour construire une référence, mais très souvent ces pratiques sont insuffisamment renseignées dans les études, sur leur expertise, ou sur la manière d'agréger leurs avis lorsque plusieurs experts sont sollicités pour une même étude. Dans cette optique, le travail de ce chapitre a consisté à proposer une méthode de production de référence à l'aide d'experts, sur la base des recommandations de Maier-Hein *et al.* (2018) de se rapprocher du terrain et des sociétés savantes liées à notre domaine d'application, l'imagerie de la scoliose, et du processus de production de modélisation 3D de la compagnie EOS Imaging. Cette nouvelle méthode s'appuie sur un consensus, comme recommandé par Cardoso *et al.* (2014) pour la construction d'une réalité terrain, et déjà avancé sans preuves ou analyses par Warfield *et al.* (2004) avec « l'expert parfait ». Nous allons analyser cette nouvelle méthode dans ce qu'elle apporte en confiance dans les mesures cliniques issues des reconstructions 3D de colonne vertébrale.

3.1.1 Les pratiques de terrain pour la reconstruction 3D de colonne

Sur le terrain, les reconstructions 3D EOS sont utilisées en clinique, avec une approbation FDA[1] et un marquage CE[2] du logiciel commercial de reconstruction démontrant la précision et la fiabilité des mesures cliniques extraites des modèles 3D. La compagnie EOS Imaging a

[1] Food and Drug Administration : administration américaine des denrées alimentaires et des médicaments

[2] Conformité Européenne

mis en place un processus (figure 3.1) constituant son modèle d'affaire, avec la création d'un département dédié à la reconstruction 3D de colonne vertébrale. Ceci est rendu possible par la disponibilité des experts, ce qui n'est pas le cas sur le terrain clinique, en institution hospitalière ou en cabinet de radiologie. Ces experts sont en effet formés à la méthode de reconstruction et l'utilisent au quotidien. Pourtant, ce processus n'a pas été éprouvé quant à son efficacité, bien que par intuition il semble naturel que plusieurs experts se mettant d'accord ensemble soit plus pertinent qu'un seul expert décidant seul, ou qu'une simple moyenne de résultats d'experts qui ne se sont pas consultés. En pratique, un premier expert (OP1) effectue la modélisation (Étape 1 : Reconstruction), puis un second (OP2) revoit et remodifie les ajustements des projections 2D du modèle 3D sur les images s'il l'estime nécessaire (Étape 2 : Revue) et renvoie cette modélisation corrigée au premier expert afin qu'il donne son accord ou son désaccord sur les modifications apportées (Étape 3 :

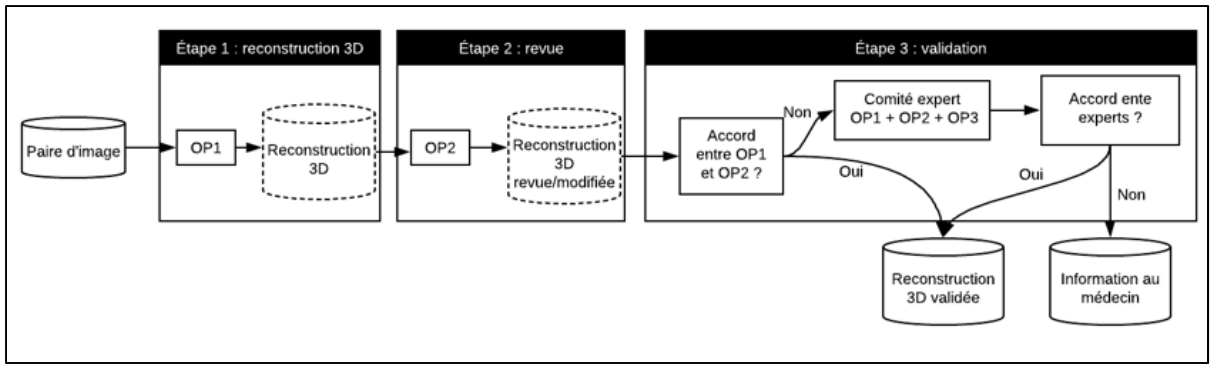


Figure 3.1 Processus de production de reconstructions 3D de l'entreprise partenaire avec les 3 étapes : Reconstruction, Revue et Validation

Validation). Si un désaccord s'engage entre les deux premiers experts, un comité expert est formé avec les deux premiers experts et un troisième expert afin de trouver un consensus sur l'ajustement des projections 2D du modèle 3D par rapport aux images. Si des doutes persistent sans que les experts ne puissent se mettre d'accord, un message personnalisé est adressé au chirurgien ou au médecin ayant demandé la modélisation, afin de l'informer des zones de désaccords sur la modélisation 3D.

3.1.2 Objectifs

Une réalité terrain est une analyse ayant fait le plus haut niveau de preuve de confiance dans les mesures. Pour cela, l'étude expérimentale présentée dans ce chapitre consistera à proposer une nouvelle méthode de création de référence inspirée du processus mis en place par EOS Imaging pour la modélisation 3D de colonne vertébrale. Pour évaluer ce processus, nous allons analyser le niveau de confiance dans les mesures cliniques apporté par l'inclusion de discussions entre experts. Le but ici est de caractériser le comportement du bassin d'experts mobilisés, au regard de la qualité des images et selon les étapes du processus, en analysant l'évolution des intervalles de confiance de mesures cliniques avant et après la réalisation du processus.

3.2 Étude de reproductibilité

3.2.1 Données

Nous décrivons dans cette section les données que nous avons utilisées pour notre analyse expérimentale.

3.2.1.1 Type de données

Soixante-quatre patients ont été inclus dans cette étude, soit 29 adultes et 35 adolescents. À chaque patient est lié une paire de radiographies bi-planes EOS ainsi que plusieurs modélisations 3D selon l'étape du processus de reconstruction auxquelles elles correspondent. Les critères d'inclusion étaient les suivants :

- Radiographies bi-planes disponibles avec acquisition des vertèbres allant de T1 à L5 et visibilité des hanches sur les deux vues;
- Sujets non opérés ;
- Reconstructions 3D effectuées sous une version commerciale de SterEOS®.

Au total, 1090 modélisations 3D sont récupérées (voir figure 3.2), parmi lesquelles :

- 192 modélisations initiales issues de l'étape d'analyse : 64 reconstructions 3D réalisées 1 fois par 3 opérateurs différents;
- 576 modélisations revues/corrigées issues de l'étape 2 de revue : 192 modélisations initiales revues/corrigées 1 fois à l'aveugle par chacun des 3 correcteurs différents dont 254 validées d'emblée après revue, c'est-à-dire sans corrections;
- 322 modélisations validées issues de l'étape 3 de validation après corrections dont 10 issues d'un consensus d'un comité de 3 experts. Les corrections ont été validées à l'aveugle, c'est-à-dire que l'expert à la validation ne sait pas qui a réalisé la modélisation initiale ni qui a apporté des corrections.

Ces données anonymisées nous ont été fournies par EOS Imaging et s'insèrent dans le certificat d'éthique *CE14.368-Projet_EOS : Projet EOS* entre EOS Imaging et le Laboratoire de recherche en Imagerie et Orthopédie de Montréal.

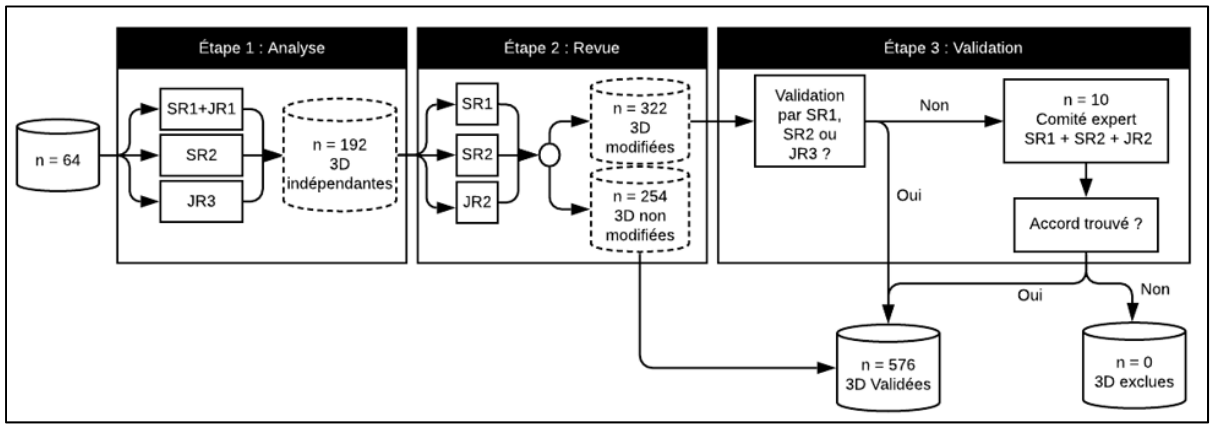


Figure 3.2 Adaptation du processus de production de modélisations 3D de l'entreprise partenaire et quantification des données partagées pour notre étude de reproductibilité

3.2.1.2 Caractéristiques des données

Afin de caractériser les données, différentes catégories ont été définies pour la base de données selon l'âge des sujets, la gravité des déformations et la visibilité des structures. Cela permet de décrire l'échantillon de données, de donner des informations sur sa représentativité

de la population, et d'affiner l'interprétation des comportements des experts et de l'évolution de variabilité des mesures.

Pour la sévérité des déformations, nous nous sommes appuyés sur le travail de Negrini *et al.* (2018) pour sélectionner la valeur d'angle de Cobb (section 1.2.2.2) qui nous permettrait de séparer les patients sévères des patients avec déformations légères ou modérées. Pour la visibilité des structures, nous nous sommes inspirés de Deschênes *et al.* (2010) pour mettre en place un score de visibilité des structures dans les images.

Pour l'âge : Les patients pédiatriques sont définis par un âge inférieur à 18 ans et les adultes sont définis par un âge égal ou supérieur à 18 ans.

Pour la sévérité : Basée sur l'angle de Cobb, 44 patients parmi les 64 ont une déformation de la colonne vertébrale sévère (au moins un angle de Cobb >40°) et 20 sujets présentent une déformation rachidienne légère à modérée (aucun angle de Cobb >40°). Ces critères sont basés sur les recommandations de Negrini *et al.* (2018) pour la prise en charge de la scoliose adolescente.

Pour la visibilité des structures : Une évaluation subjective de la visibilité a été réalisée en amont des analyses par un expert en anatomie, en attribuant une note de 0, 0,5 ou 1 à chaque structure d'intérêt de chaque vertèbre (respectivement : pas du tout visible, partiellement visible, et totalement visible). Un score normalisé de 0 à 100 a ensuite été calculé pour chaque patient (moins de 50 = mauvaise visibilité ; plus de 50 = bonne visibilité, pour la colonne entière, pour des régions de la colonne, et par vue radiographique. Ces critères de visibilité des structures sont inspirés des travaux de Deschênes *et al.* (2010).

Les structures d'intérêts sont les pédicules en vue frontale, les plateaux en vue frontale et en vue latérale, la pente sacrée en vue latérale (voir figure 3.3), car ce sont les structures prises comme repères prioritaires par les opérateurs (section 1.2.2.1, figure 1.4).

L'attribution des notes correspond pour les plateaux et la pente sacrée au type de visualisation de ces structures : le plateau apparaît similaire à une ligne car tangent au rayon X (complétement visible) ; le plateau apparaît dédoublé ou ellipsoïdal car non tangent au rayon X (partiellement visible) ; le plateau est indéfinissable (pas du tout visible) (voir figure

3.4). L'attribution des notes correspond pour les pédicules à une adaptation de la classification de Nash Moe (O'Brien & Spinal Deformity Study Group, 2004) de rotation vertébrale, selon la visibilité des pédicules (2 pédicules symétriques et visibles = complètement visible ; pédicules asymétriques mais totalement ou partiellement visibles = partiellement visible ; un seul pédicule visible = pas du tout visible) (voir figure 3.5).

L'annexe I synthétise des caractéristiques sur l'échantillon de sujets ainsi que la taille des sous-échantillons au regard de ces critères.

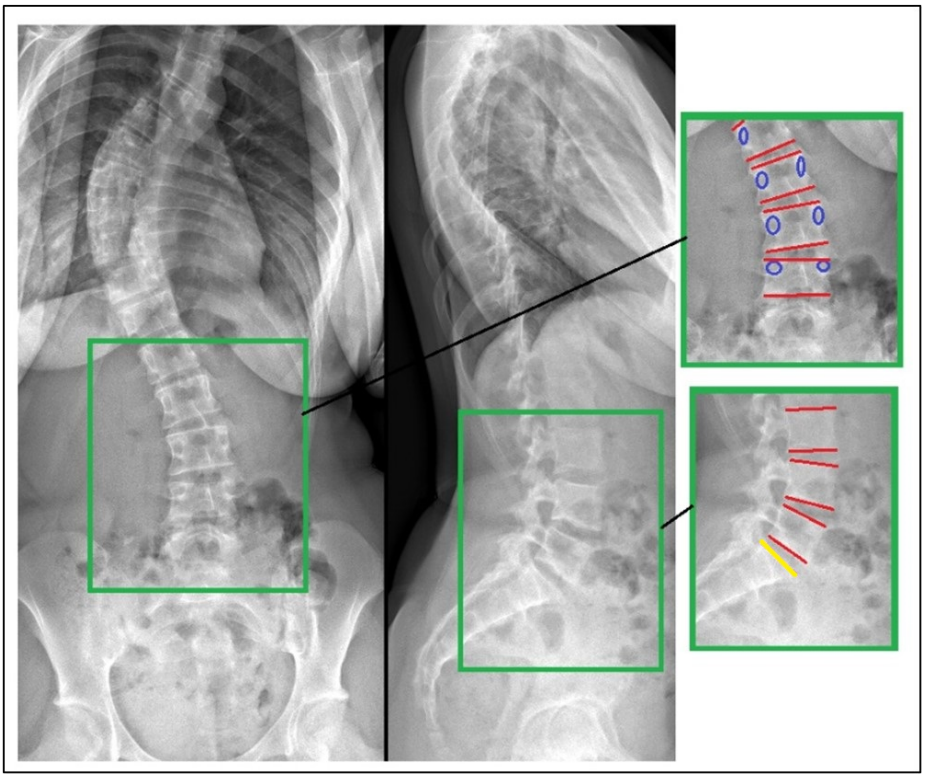


Figure 3.3 Structures d'intérêts de la colonne vertébrale, avec agrandissement de la région lombaire ; les pédicules sont visibles en bleu, les plateaux en rouge, et la pente sacrée en jaune

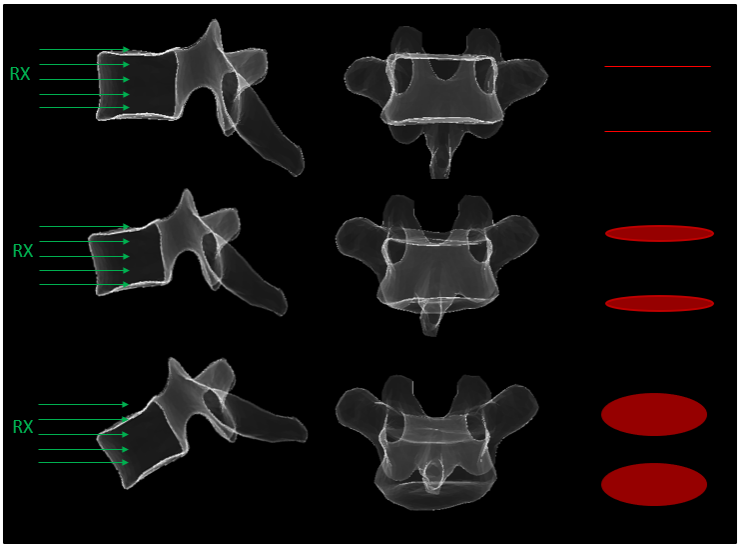


Figure 3.4 Exemple de visualisation des plateaux par rapport à l'orientation des rayons X et de la vertèbre ; des rayons tangents permettent la visualisation de plateaux sous la forme d'une ligne (en haut); des rayons non tangents provoquent un dédoublement visuel des plateaux sous la forme d'une ellipse (au milieu); lorsque le dédoublement est trop accentué, les contours deviennent difficilement identifiables (en bas)

3.2.2 Experts mobilisés

Cinq experts de l'entreprise partenaire ont été recrutés pour cette étude, avec des expériences de modélisations différentes :

- 4 experts différents pour l'étape de reconstructions 3D initiales dont trois avec 1 an (JR1), 2.5 ans (SR2), et 4 ans (SR1) d'expérience en modélisation à temps plein, et un (JR3) avec 1 an d'expérience en modélisation à temps partiel. Chacun des experts a réalisé les reconstructions 3D pour les 64 sujets, à l'exception de SR1 et JR1 qui se sont partagés respectivement les adultes et les adolescents pour des raisons logistiques liées aux activités de l'entreprise.
- 3 experts différents ont participé à l'étape de revue, SR1 et SR2 (qui ont également participé aux modélisations initiales), et JR2, qui dispose de 1 an d'expérience de modélisation à temps plein. Les corrections se sont effectuées en aveugle, c'est-à-dire qu'un correcteur ne savait pas qui il corrigeait, afin de limiter le biais du senior.

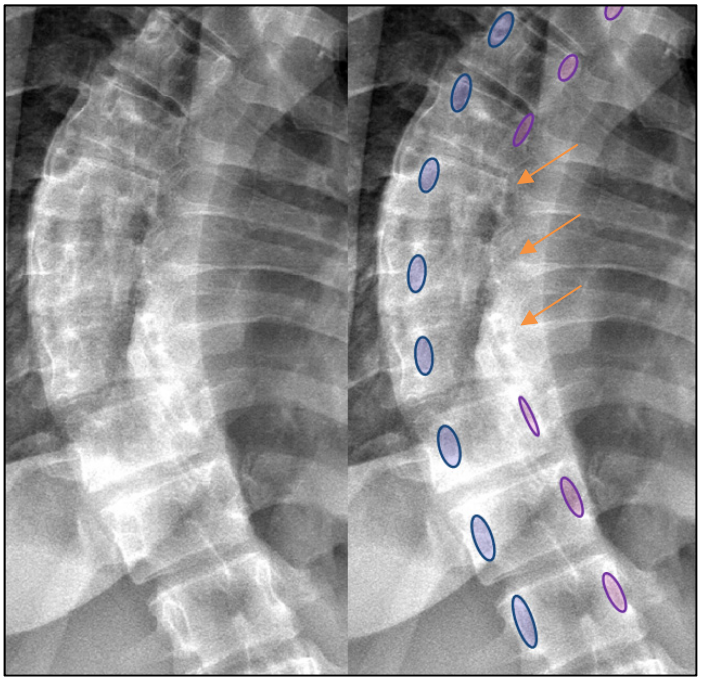


Figure 3.5 Exemple de visualisation des pédicules : les pédicules à la droite du patient (en bleu) sont tous visibles sur cette partie de la colonne, alors que les pédicules gauche (en mauve) disparaissent dans l'image (flèches) à cause des déformations osseuses

Aussi, les corrections pour un même patient se sont faites à plusieurs jours d'intervalles pour limiter le biais de mémoire. Les auto-corrections ont concerné 279 des 576 modélisations revues.

- 3 experts se sont partagé les modélisations corrigées à valider pour l'étape de validation, et ont participé au comité expert en cas de désaccord : SR1, SR2 et JR2.

Ces experts sont tous formés en radio-anatomie depuis plusieurs années, avec chacun un diplôme de technologue en radiologie médicale.

3.2.3 Mesures cliniques

Les mesures cliniques retenues pour l'étude sont celles extraites automatiquement par le logiciel de reconstruction SterEOS® depuis la modélisation 3D (voir figure 3.6) : Angle de Cobb, Cyphoses T1/T12 et T4/T12, Lordoses L1/L5 et L1/S1, inclinaison du pelvis PT, incidence pelvienne PI, pente sacrée SS, rotations axiales, frontales et latérales des vertèbres de T1 à L5.

3.2.4 Analyses statistiques

Nous voulons observer la variabilité des mesures cliniques selon l'étape du processus décrit en figure 3.2, afin d'analyser l'intérêt du processus pour la construction d'une référence, afin d'identifier s'il permet d'avoir un haut niveau de confiance dans les mesures cliniques et de construire une réalité terrain.

Les analyses statistiques ont été réalisées sur StatGraph®. Pour chaque étape du processus de modélisation, la moyenne et l'écart type de chaque mesure clinique par patient ont été calculés afin d'analyser la variation de la moyenne des mesures selon l'étape du processus et de calculer des intervalles de confiance des mesures cliniques d'autre part.

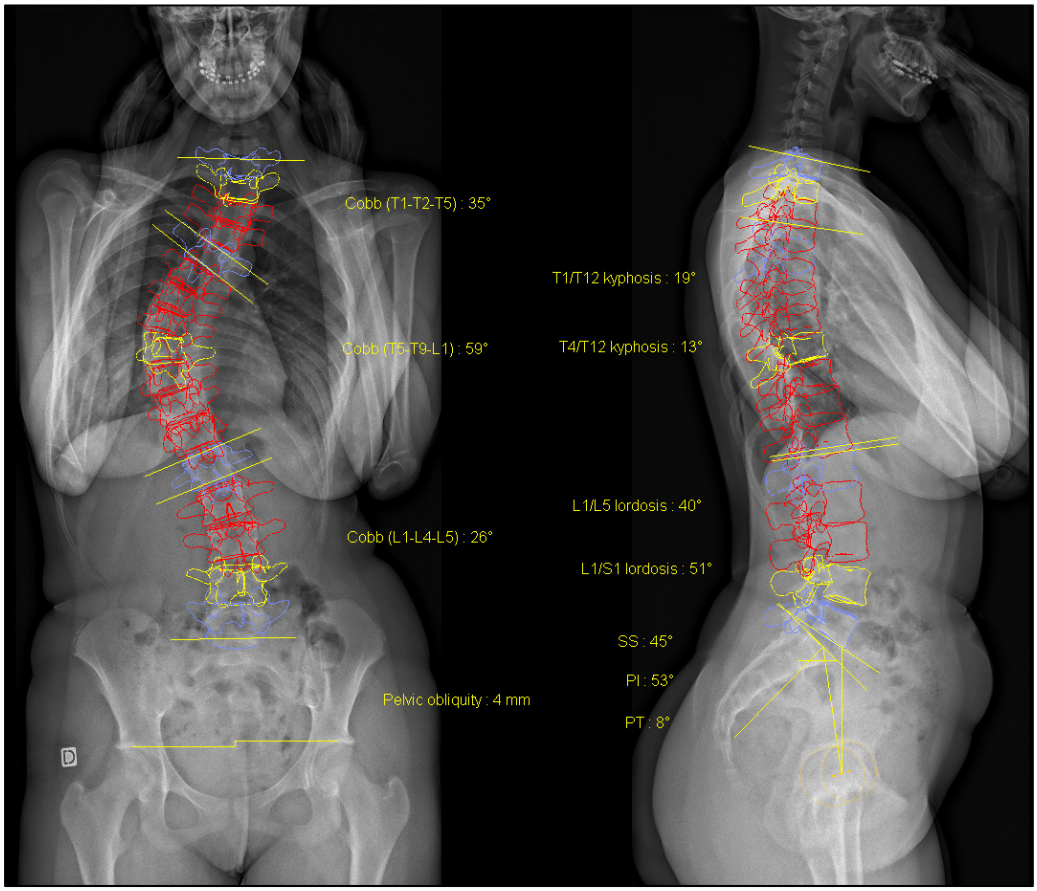


Figure 3.6 Paramètres cliniques en sortie du logiciel commercial de reconstruction 3D de la colonne vertébrale

Les intervalles de confiance à 95% ont été calculés selon la formule de Glüer *et al.* (1995) et les recommandations de la norme ISO 5725-1 (équation 1.14). Les valeurs de ±2RMSSD ont été calculés pour l'échantillon complet mais aussi par catégorie de patients (sévérité, âge, visibilité de région et de vue) afin d'analyser la reproductibilité des mesures cliniques selon l'étape de reconstruction 3D et selon les types d'images rencontrés. La variation de reproductibilité (VR) est calculée comme le pourcentage de variation depuis la reproductibilité initiale (équation 3.1). Si la variation de reproductibilité est positive, il s'agit d'un gain de reproductibilité.

$$VR = \frac{RMSSD_{validation} - RMSSD_{analyse}}{RMSSD_{validation}} \times 100 \quad (3.1)$$

3.3 Résultats

3.3.1 Comportement du bassin d'experts

Parmi toutes les corrections apportées aux modélisations, on observe que l'immense majorité de ces corrections se sont faites pendant l'étape de revue (97% des retouches contre 3% seulement à l'étape de validation) (voir tableau 3.1). Les modifications ont été classées par les correcteurs selon l'amplitude des impacts sur les mesures cliniques. Une modélisation non retouchée était classée conforme, une modélisation retouchée mais avec peu d'effet sur les mesures cliniques était classée non conforme avec modification mineure, et une modélisation retouchée avec un impact de plus de 3° ou 3 mm sur les mesures cliniques était classée non conforme. Un test du χ^2 a démontré que les sujets adultes ont eu significativement plus de modélisations jugées non conformes que de modélisations conformes (valeur de p = 0,0201) et c'est la seule catégorie de patient pour laquelle un taux de significativité a été atteint pour ce test.

Tableau 3.1 Répartition du nombre de modifications apportées entre l'étape de Revue et l'étape de Validation.

Étape du processus	REVUE	VALIDATION	Total
Nombre de corrections	322	10	332
Pourcentage de corrections	97%	3%	100%

Nous avons regardé la répartition des retouches selon l'expert ayant produit la modélisation à l'étape de reconstructions 3D et selon l'expert correcteur à l'étape de revue, afin d'analyser le comportement du bassin des experts, sachant que les corrections ont été faites à l'aveugle. Le tableau 3.2 montre la répartition des types de corrections selon l'expert corrigé. On voit que l'expert le moins expérimenté et l'expert le plus expérimenté (JR3 et SR1) ont reçu le plus de corrections non conformes. Les données de SR1 ne concernent cependant que les adultes. De la même façon, les données de JR1 ne concernent que les adolescents, et on observe qu'il s'agit de l'expert avec le moins de corrections non conformes. Si l'on combine ces deux experts pour avoir l'échantillon complet, on obtient un score similaire à SR2.

Tableau 3.2 Répartition des corrections par modélisateur initial.

Quand on est corrigé	**SR2**	**JR3**	**JR1**	**SR1**	**JR1/SR1**
Conformes	50%	35%	54%	39%	47%
Non Conformes - Modification mineure	9%	7%	18%	1%	12%
Non conformes	41%	58%	28%	60%	41%

Le tableau 3.3 s'intéresse au comportement des correcteurs. On voit que les comportements de chacun est similaire au regard du même échantillon de 192 modélisations initiales. On note que le correcteur le plus expérimenté (SR1) a apporté davantage de corrections non conformes que les autres correcteurs.

Tableau 3.3 Répartition des corrections en fonction de l'expert correcteur.

Quand on est correcteur	**SR2**	**JR2**	**SR1**
Conformes	46%	45%	41%
Non Conformes - Modification mineure	10%	10%	9%
Non conformes	44%	45%	50%

Le tableau 3.4 synthétise la répartition des modifications, mineures ou non, selon le correcteur et le modélisateur initial (Lire : lorsque SR2 a corrigé SR2, il a identifié 39% de

modélisations non conformes). On peut voir que JR3 est l'expert ayant reçu le plus de modifications, tous correcteurs confondus. SR1 semble avoir reçu plus de corrections que JR1. Concernant les auto-corrections, nous pouvons voir que SR2 a estimé non conformes 39% de ses propres modélisations, et SR1 a estimé non conformes 59% de ses propres modélisations.

Tableau 3.4 Répartition des modélisations non conformes selon le correcteur et le modélisateur initial.

Correcteur / Modélisateur	SR2	SR1	JR2
SR2	39%	52%	42%
SR1	69%	59%	55%
JR1	40%	49%	46%
JR3	69%	73%	55%

3.3.2 Variabilité des mesures cliniques

Sur l'échantillon de 64 patients, il n'existe pas de différences significatives (ANOVA) entre les moyennes des paramètres cliniques, que l'on regarde les moyennes à l'étape de reconstructions 3D (avec 3 mesures initiales), à l'étape de revue (avec 3 mesures corrigées) ou à l'étape de validation (avec 9 mesures validées). La figure 3.7 montre les boîtes à moustache des mesures sagittales pour les modélisations de l'étape d'analyse et de l'étape de validation. Pour les mêmes mesures, il n'existe pas de différences significatives (ANOVA) entre les écarts types des paramètres cliniques observés à l'étape de reconstructions 3D et ceux observés à l'étape de validation. On peut cependant observer sur les boîtes à moustache une diminution systématique des écarts types à l'étape de validation. Ces résultats ont été également constatés pour les autres paramètres cliniques.

3.3.3 Variation de la reproductibilité

Les mesures de 2RMSSD inter-opérateur et intra-correcteurs pour les mesures cliniques classiques sont synthétisées dans la figure 3.8. On peut voir que le 2RMSSD pour les paramètres pelviens est plus élevé sur les mesures inter-opérateurs (3 mesures initiales) que sur les mesures intra-correcteurs (3 mesures corrigées par JR2, SR2 ou SR1).

La figure 3.8 synthétise l'évolution de la grandeur des intervalles de confiance en pourcentage selon les étapes du processus de modélisation considérées, par catégories de paramètres cliniques. L'évolution est positive pour tous les paramètres cliniques, ce qui veut dire que les pourcentages affichés constituent tous un gain de reproductibilité, soit un resserrement des intervalles de confiance. On observe que le gain en reproductibilité est nettement supérieur en sortie d'étape de validation qu'à l'étape de revue (25,7% d'amélioration contre 15,5-18,5% pour les paramètres pelviens).

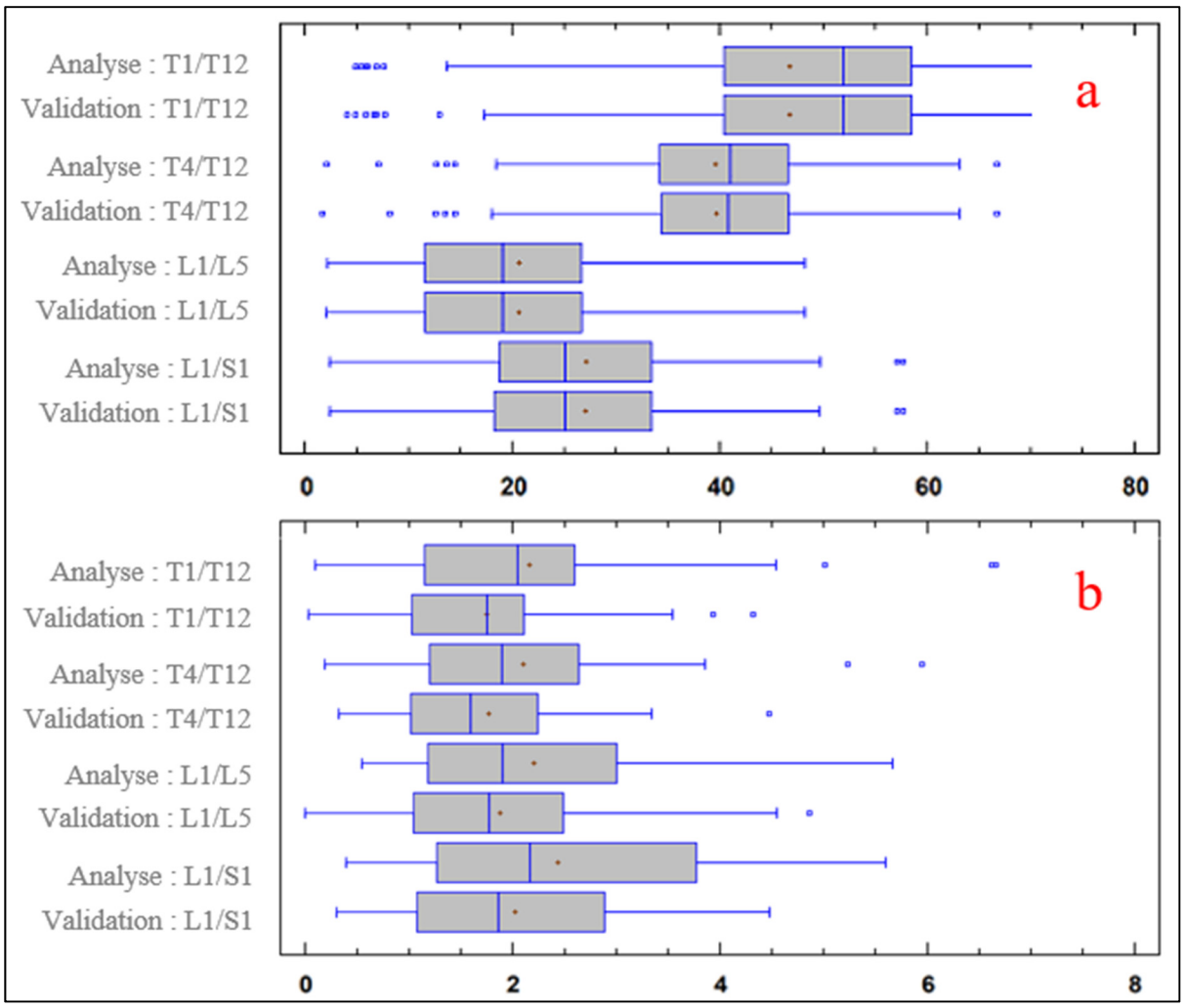


Figure 3.7 Boîtes à moustache des moyennes (a) et écarts types (b) des paramètres de cyphoses et de lordoses pour les modélisations de l'étape d'analyse et les modélisations de l'étape de validation

Le tableau 3.5 nous communique aussi des informations sur le comportement des experts correcteurs, qui contribuent tous à l'augmentation de reproductibilité en sortie de l'étape de validation. SR1 augmente davantage la reproductibilité des mesures pelviennes (18,5%) que les autres correcteurs, et pour la balance sagittale c'est SR2 qui accorde plus de confiance dans ses mesures (10,3% d'augmentation contre 4,6% pour SR1 et 7,2% pour JR2).

Les figures 3.9 et 3.10 montrent le gain en reproductibilité par catégories de mesures cliniques selon certaines catégories de patients. La figure 3.9 montre que l'amélioration de reproductibilité est supérieure pour les patients avec une mauvaise visibilité sur la vue frontale pour les paramètres pelviens, les paramètres de balance sagittale et les paramètres de scoliose (respectivement 24%, 23%, 19% contre 12%, 13,5% et 11% pour les patients avec une bonne visibilité frontale). De même, l'amélioration en reproductibilité est nettement supérieure pour les patients avec une mauvaise visibilité des lombaires, pour les paramètres pelviens, les paramètres de balance sagittale et les paramètres de scolioses (respectivement 24%, 23 %, 16,5% contre 12%, 13,5% et 11,5% pour les patients avec une bonne visibilité lombaire). En revanche, les différences d'augmentation de reproductibilité sont moins perceptibles pour les catégories d'âges (figure 3.10).

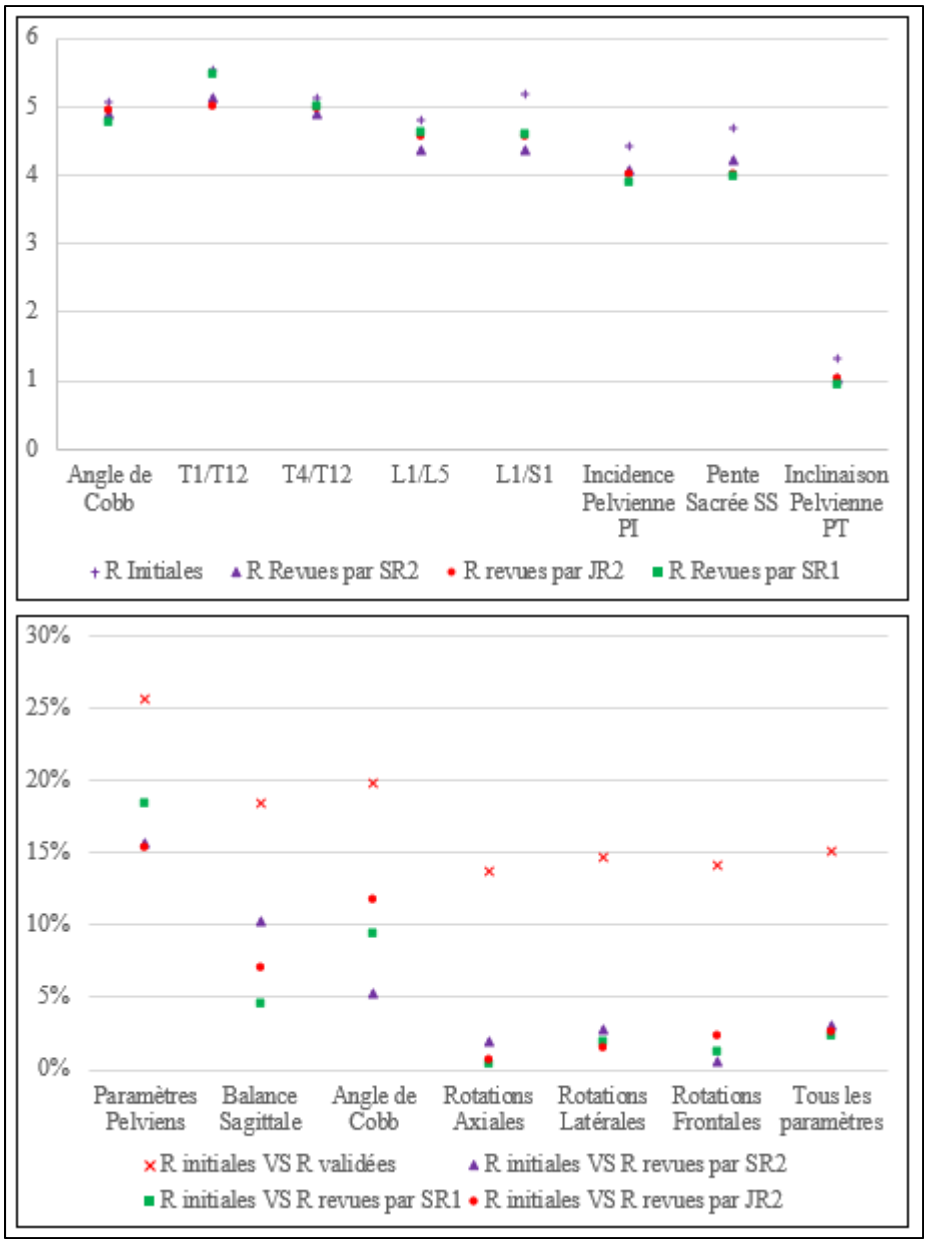


Figure 3.8 En haut : Reproductibilité (2RMSSD) inter-opérateur et intra-correcteurs pour l'angle de Cobb, les mesures de cyphoses et de lordoses ainsi que les paramètres pelviens. En bas : Variation de reproductibilité (%2RMSSD) selon différentes catégories de paramètres cliniques et selon le correcteur

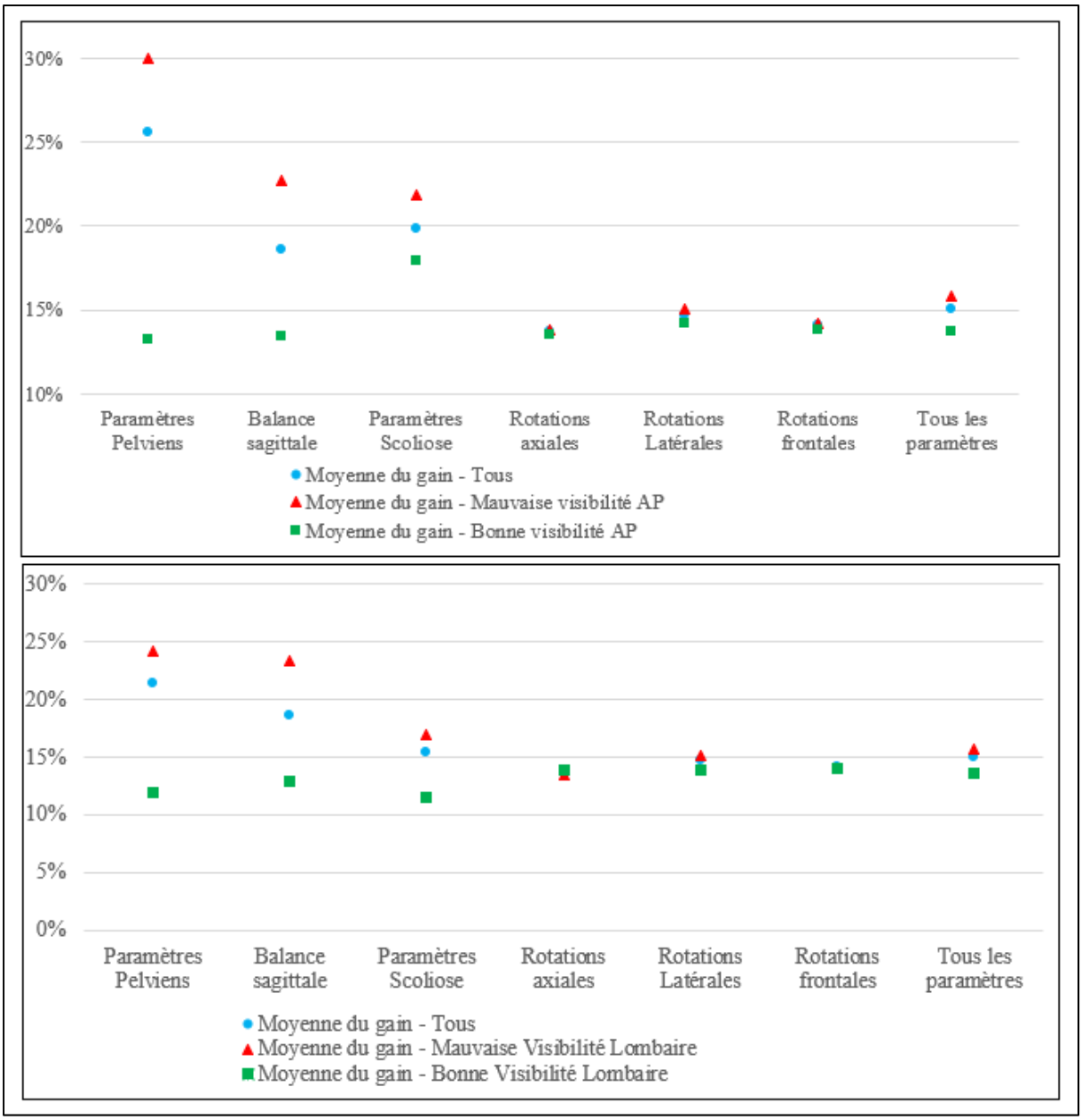


Figure 3.9 Variation de reproductibilité entre les mesures cliniques issues des reconstructions 3D initiales et celles issues des reconstructions 3D validées, pour différentes catégories de paramètres cliniques et pour différentes qualités de visibilité (en haut : visibilité en vue frontale ; en bas : visibilité des structures lombaires)

Tableau 3.5 Variation de reproductibilité (VR) moyenne (et écart type) selon les catégories de paramètres cliniques et selon l'étape du processus observée

VR Moyenne	Paramètres Pelviens	Balance sagittale	Scoliose	Rotations axiales	Rotations Latérales	Rotations frontales	Tous
Entrée VS Sortie	**25,7%**	**18,6%**	**15,4%**	**13,7%**	**14,7%**	**14,1%**	**15,0%**
Écart type	7,1%	3,3%	4,9%	2,0%	1,6%	1,1%	3,7%
Entrée VS SR2	**15,6%**	**10,3%**	**2,6%**	**2,0%**	**2,8%**	**0,6%**	**3,0%**
Écart type	10,3%	6,0%	4,1%	5,2%	4,1%	2,7%	5,6%
Entrée VS SR1	**18,5%**	**4,6%**	**5,1%**	**0,5%**	**2,0%**	**1,3%**	**2,6%**
Écart type	9,0%	4,4%	5,8%	3,8%	1,9%	1,2%	5,1%
Entrée VS JR2	**15,5%**	**7,2%**	**3,9%**	**0,8%**	**1,5%**	**2,4%**	**2,6%**
Écart type	7,1%	4,2%	5,5%	2,4%	2,0%	2,2%	4,6%

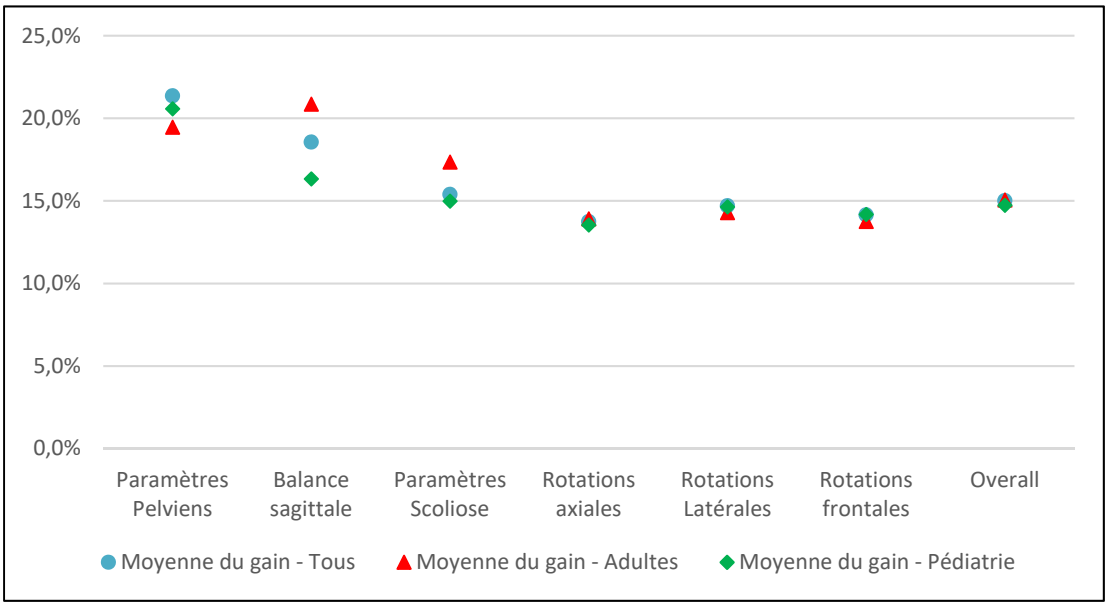


Figure 3.10 Gain en reproductibilité apporté par chaque correcteur pour différentes catégories de paramètres cliniques, pour les adultes et pour les adolescents

Tableau 3.6 Reproductibilité des paramètres cliniques selon la visibilité en vue frontale et comparaison des valeurs à la littérature ; les données d'entrée sont les 3 modélisations initiales indépendantes de l'étape de reconstructions 3D, les données de sortie sont les 9 modélisations validées à la suite du processus de l'entreprise partenaire

	Tous patients		**Bonne visibilité AP**		**Mauvaise visibilité AP**		**Sévères**	
2RMSSD (°)	Entrée n = 64	**Sortie n = 64**	Entrée n = 37	**Sortie n = 37**	Entrée n = 27	**Sortie n = 27**	**Humbert 2009**	**Ilharreborde 2011**
Cobb1	4,9	**3,9**	4,7	**3,8**	5,1	**4,0**	3,5	6,2
Cobb2	3,7	**3,1**	3,8	**3,1**	3,5	**3,1**		
Cobb3	5,6	**4,4**	4,8	**4,1**	6,6	**4,8**		
T1/T12	5,6	**4,6**	5,3	**4,3**	5,9	**4,9**	5,6	7,0
T4/T12	5,1	**4,3**	4,6	**4,1**	5,7	**4,6**	4,3	5,7
L1/L5	4,8	**4,0**	4,1	**3,6**	5,7	**4,4**	5,4	6,7
L1/S1	5,2	**4,0**	4,1	**3,6**	6,5	**4,5**	4,2	5,9
PI	4,4	**3,6**	3,4	**2,9**	5,6	**4,3**	3,5	4,7
SS	4,7	**3,6**	3,3	**2,8**	6,1	**4,5**	3,2	4,3
PT	1,3	**0,9**	0,8	**0,7**	1,8	**1,1**	0,8	1,4

Le tableau 3.6 résume les différences de reproductibilité entre les échantillons de patients avec mauvaise visibilité frontale, les patients avec une bonne visibilité frontale, l'échantillon complet de patients et enfin les résultats de deux études similaires de reproductibilités des mesures cliniques pour des patients sévères (Humbert *et al.*, 2009; Ilharreborde *et al.*, 2011). On voit d'abord que le processus de modélisation de l'entreprise EOS Imaging permet de retrouver des intervalles de confiance pour les patients à mauvaise visibilité plus proche des intervalles de confiance pour les patients avec une bonne visibilité (différence moyenne de 1,42° entre les reproductibilités d'entrée contre seulement 0,72° de différence moyenne entre les reproductibilités de sortie). Si l'on se compare aux données de la littérature, les reproductibilités que l'on obtient sur les mesures initiales sont comparables ou meilleures que celles obtenues par Ilharreborde *et al.* (2011). Les reproductibilités que l'on obtient sont similaires ou légèrement moins bonnes que celles de Humbert *et al.* (2009), mais si l'on

regarde l'échantillon de patients avec une bonne visibilité, les reproductibilités sont meilleures pour les cyphoses, lordoses et paramètres pelviens sur les modélisations validées.

3.4 Discussion

Notre étude consistait à proposer une méthode de construction de référence en analysant un processus de validation de reconstructions 3D mis en place par EOS Imaging qui repose sur des discussions entre experts et un consensus. Le but est d'identifier les impacts des corrections et des discussions sur les mesures cliniques et leurs fiabilités, pour savoir si les données validées en sortie du processus présentent un plus haut niveau de confiance que les données issues d'experts indépendants. Cela devait nous permettre de conforter l'idée de « l'expert parfait » avancée mais non démontrée par Warfield *et al.* (2004) et de valider le processus de reconstruction de l'entreprise partenaire comme processus permettant d'obtenir une haute confiance dans les mesures cliniques.

L'analyse du comportement d'expert nous permet d'observer plusieurs points :

- Nous avons sollicité des experts formés depuis plus d'un an sur le logiciel de modélisation, et depuis plusieurs années en radio anatomie. Or nous avons vu (voir tableau 3.2) que même l'expert avec le plus d'expérience a reçu un nombre de corrections importants sur ses modélisations (60% de modélisations non conformes), quasi totalement acceptée pour la validation. Ce résultat est conforté par la mise en place des corrections à l'aveugle, ce qui a limité le biais du senior (un expert junior aurait peut-être donné moins de modifications s'il savait qu'il corrigeait la modélisation d'un senior).
- La grande majorité des modifications se sont faites à l'étape de revue (voir tableau 3.1), avec un resserrement des intervalles de confiances observables sur tous les paramètres cliniques (voir tableau 3.5). Les experts sollicités sont en effet formés par l'entreprise partenaire pour modéliser et identifier les points à corriger pour augmenter la confiance dans les mesures, ce qui les rend efficaces dès l'étape de revue.

- Les corrections apportées se sont majoritairement concentrées en intensité sur les paramètres pelviens, les cyphoses et lordoses, l'angle de Cobb, et moins sur les rotations vertébrales (voir figure 3.8).
- Il existe des différences de corrections entre les experts : certains experts vont avoir plus de confiance dans certains paramètres que dans d'autres (voir figure 3.8), bien que les 3 modélisations initiales aient été corrigées à plusieurs jours d'intervalles pour limiter le biais de mémoire.
- Si l'échantillon adulte semble recueillir beaucoup plus de modifications que l'échantillon adolescent (voir tableau 3.2), cela ne semble pas se refléter sur l'augmentation de reproductibilité (voir figure 3.10). Cela veut dire que l'amplitude des modifications apportées compense la fréquence des modifications, ou que le peu de modifications apportées à l'échantillon adolescent suffit à augmenter la reproductibilité au moins autant que pour l'échantillon adulte.
- Il existe de nombreuses auto-corrections (voir tableau 3.4) avec 59% de modélisations non conformes déclarées par SR1 pour SR1 et 39% par SR2 pour SR2). Cela peut être dû à des facteurs humains liés à la complexité de la tâche (fatigue, habitudes, oublis) et renforce la nécessité de contre-vérifications par un expert quel qu'il soit.

Sur l'évolution de la reproductibilité des mesures, nous observons ces points :

- Malgré la non-significativité des différences de dispersion (écart-types) entre les modélisations initiales et les modélisations validées, des analyses plus poussées sur l'évolution de reproductibilité nous montrent un resserrement systématique des intervalles de confiance, pour toutes les mesures cliniques.
- L'augmentation de reproductibilité n'est pas uniquement dû au plus gros nombre de modélisations incluses dans le calcul (3 modélisations initiales contre 9 modélisations validées) (voir figure 3.8). On observe en effet que parmi les correcteurs, chacun permet une augmentation de la reproductibilité (3 modélisations initiales contre 3 modélisations corrigées).
- Il existe une influence de la visibilité image sur la reproductibilité des mesures (voir figure 3.9). Pour la majorité des mesures cliniques, la reproductibilité est d'emblée

moins bonne pour les patients avec une mauvaise visibilité sur les modélisations initiales (voir tableau 3.5). Pour les paramètres pelviens, elle est presque deux fois moins bonne (6,1 degrés contre 3,3 degrés pour la pente sacrée). Le score de visibilité ayant été construit selon la visibilité des structures, cela peut être dû à de grosses déformations empêchant la correcte visualisation des structures comme cela peut être dû à une qualité d'image dégradée.

- La différence de reproductibilités identifiée entre l'étude de Ilharreborde *et al.* (2011) et la nôtre (tableau 3.6) peut s'expliquer par des experts mieux formés, par l'échantillon de patients plus grand et l'inclusion de patients adultes avec des déformations légères ou modérées. Pour les angles de Cobb, les deux études citées ne tiennent compte que d'un seul angle de Cobb, sans préciser comment il a été sélectionné. Les niveaux vertébraux étaient présélectionnés dans les études citées tandis que dans notre analyse, si les niveaux ont été présélectionnés à l'étape de reconstructions 3D, les experts avaient le droit de les modifier pendant les étapes de corrections. Pourtant, en sortie de processus, sur les modélisations validées, nos valeurs de reproductibilités sont améliorées en dépit de cette possibilité de modifications, et meilleures que celles de Ilharreborde *et al.* (2011) quel que soit l'angle observé.

3.5 Résumé

Ce premier travail nous amène de nombreuses réflexions intéressantes sur la manière de solliciter des experts lorsque l'on souhaite construire une référence. Nous l'avons constaté avec Maier-Hein *et al.* (2018), le manque de standardisation et de recommandations a entrainé des défauts d'informations dans de nombreuses méthodologies d'évaluations, et surtout une faible robustesse des résultats selon l'expert qui a construit la référence. Or, même un expert très expérimenté peut faire des erreurs, comme nous l'avons vu dans ce travail. Amener une confrontation des opinions d'expert, sous la forme de corrections discutées comme dans le processus de l'entreprise partenaire, permet en revanche d'augmenter la confiance dans les valeurs cliniques que l'on souhaite pouvoir prendre comme référence. Évidemment, dans notre étude, nous avons vu que statistiquement il n'y a

pas de différence des moyennes des mesures avant ou après le processus de traitement, ce qui semblerait aller dans le sens de ceux qui proposent une moyenne d'expert comme référence ou un vote majoritaire (Heimann *et al.*, 2009; Şerbănescu *et al.*, 2020; Yang *et al.*, 2021). Cependant, nous démontrons que ces étapes de processus nous permettent d'obtenir une référence plus fiable que si les experts ne s'étaient pas concertés, avec des intervalles de confiances des mesures cliniques qui sont plus étroits, ce qui rejoint la définition de la réalité terrain de Cardoso *et al.* (2014).

Un autre point majeur consiste en l'expérience des experts mobilisés. Nous avons vu que dans notre expérimentation, les experts mobilisés sont habitués à la méthode de reconstruction 3D et au processus de validation, avec une habitude de repérer les non-conformités depuis parfois plusieurs années. Dans une autre application ce point peut constituer un biais. En effet, la plupart des études d'analyses d'images qui mobilisent des experts les forment pendant quelques jours sur leurs méthodes d'analyses pour construire la référence (Aubert *et al.*, 2016; Ilharreborde *et al.*, 2011) alors que dans notre expérimentation, ce sont des experts chevronnés à la fois sur la radio-anatomie mais aussi sur le logiciel de reconstruction 3D. Cependant, notre étude prouve qu'il est possible même pour des experts seniors de faire des erreurs, et qu'il est possible de resserrer les intervalles de confiances dans les mesures cliniques.

De plus, une meilleure identification des particularités des données d'évaluation (visibilité image, sévérité des déformations, âge) permet non seulement de renseigner plus clairement la représentativité de ces données, elle permet aussi d'établir des intervalles de confiance par catégorie de population, et de mieux observer la variabilité de la confiance que l'on peut apporter à des mesures. Ce point rejoint les reproches qui sont faits aux références fantômes ou simulées, car ces méthodes ne sont pas assez représentatives des images rencontrées sur le terrain. S'il est en effet difficile dans une base de données d'évaluations de représenter 100% des caractéristiques de la population en sévérité, âge et qualité image, nous pensons qu'il est possible de limiter les biais d'échantillons en caractérisant correctement une base de données d'évaluation. De cette manière, les évaluations sont contextualisées et davantage reproductibles.

Nous nous sommes inspirés du processus de reconstruction 3D mis en place par l'entreprise partenaire pour valider des reconstructions 3D de colonne vertébrale. Si dans leur routine décrite à la figure 3.1 un seul expert est mobilisé dans les étapes d'analyse et de revue, nous avons montré que multiplier les experts, en l'occurrence 3 experts à la première étape et 3 experts à la deuxième, permet de combiner des « profils » d'expert qui, comme nous l'avons vu, n'ont pas tous exactement le même comportement de modélisation ou de corrections. Cela nous permet donc de construire une réalité terrain plus fiable car elle relève de l'opinion de plusieurs experts.

Nous pouvons de ce travail construire une référence pour l'évaluation de méthodes de reconstructions 3D de colonne vertébrale, sous la forme d'intervalles de confiance des mesures cliniques issues de notre méthode de création de référence. Notre réalité terrain n'est donc pas un résultat unique de reconstructions 3D, comme la référence de segmentation proposée par Warfield *et al.* (2004) et le STAPLE. Les intervalles de confiances des mesures cliniques permettent de prendre en compte la variabilité des mesures entre les experts, au regard de leurs performances et de leur capacité à identifier des erreurs de mesures lors des étapes de revue et de validation. Ces intervalles sont plus étroits, plus difficiles à atteindre, et donc plus exigeants et plus fiables, et constituent notre réalité terrain. Ce type de référence s'éloigne de ce que proposait Warfield *et al.* (2004) avec le STAPLE, puisqu'il s'agit d'un intervalle de confiance décrivant la référence et non d'une segmentation unique, ou pour notre domaine, d'une reconstruction 3D unique. Nous pouvons cependant utiliser le centre des intervalles de confiance, obtenu par la moyenne des mesures validées constituant la référence, pour pouvoir utiliser des métriques d'évaluation nécessitant une valeur unique de référence. Cette moyenne est fiable puisque l'intervalle de confiance autour d'elle est plus étroit.

Notre méthode de construction de référence consiste à faire appel à trois experts pour l'étape d'analyse, trois experts pour l'étape de revue, et trois experts pour l'étape de validation. Nous avons vu qu'il n'est pas nécessaire que les experts soient différents d'une étape à l'autre, du moment que les revues soient bien effectuées en aveugle pour limiter le biais du senior, dont nous n'avons pas investigué les effets éventuels. Au regard de l'expérience de chacun, nous

avons vu qu'un expert junior donne un même taux de corrections qu'un expert senior, et de la même façon, un expert senior reçoit autant de corrections qu'un expert junior. Combinés ensemble, tous ces profils participent à l'augmentation de reproductibilité des mesures cliniques, et donc à la production d'intervalles de confiance plus étroits.

Pour la suite nous nous pencherons sur une méthode d'évaluation des méthodes de reconstructions 3D de la colonne vertébrale qui puisse inclure cette nouvelle référence, que nous appellerons « intervalles de confiance 3DS », mais aussi des critères cliniques qui permettent de hiérarchiser les erreurs de mesures.

CHAPITRE 4

MÉTHODE D'ÉVALUATION

4.1 Introduction

Dans ce chapitre nous mettons en place une nouvelle méthode d'évaluation qui inclut les aspects cliniques liés à la reconstruction 3D de la colonne vertébrale avec le système EOS. Cette méthode d'évaluation inclut 4 composantes:

- Des références le plus proche possible d'un étalon or
- Des données images représentatives de la réalité
- Des métriques d'efficacité pertinences
- Un outil permettant de visualiser et exporter les résultats

Nous disposons d'une réalité terrain à la suite des travaux du chapitre 3, construite sur une base de données de 64 patients dont nous avons également détaillé la représentativité en termes d'âge, de sévérité et de visibilité des structures. Les métriques sélectionnées dans notre méthode d'évaluation correspondent aux métriques permettant d'évaluer la performance de mesures dans les images (section 1.4.4). Concernant les critères avec lesquels nous allons qualifier les erreurs de reconstructions 3D, nous allons nous rapprocher des sociétés savantes liées à la prise en charge de pathologies de la colonne vertébrale, afin de proposer des critères cliniques permettant de hiérarchiser les erreurs selon la gravité de leurs impacts potentiels sur la prise en charge d'un patient. Pour cela, nous nous appuierons sur des classifications de pathologies issues de la littérature médicale, pour identifier notamment des seuils de mesures cliniques importants. Enfin, notre méthode inclura un outil pour naviguer dans les résultats en offrant des visuels hiérarchisés par importance clinique, sous la forme de rapports d'évaluation (voir figure 4.1).

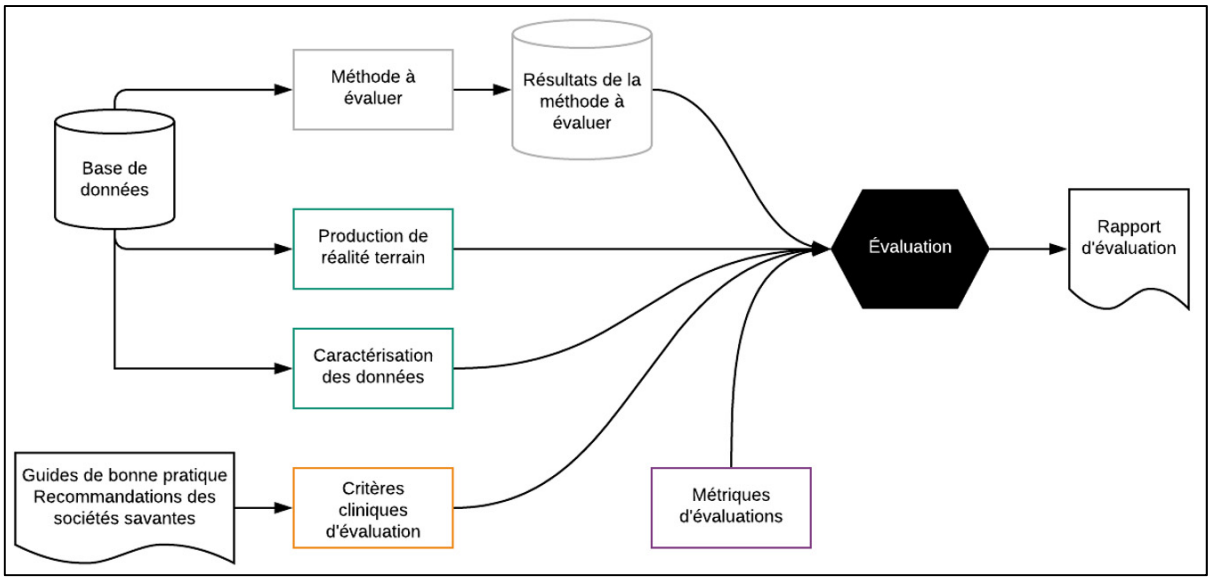


Figure 4.1 Schéma de la méthode d'évaluation générale avec les trois axes d'évaluations (Encadré vert : Construction de la réalité terrain et caractérisation des données d'évaluations ; Encadré jaune : Qualification des erreurs par élaboration des critères cliniques d'évaluations à l'aide des recommandations des sociétés savantes ; Encadré mauve : Quantification des erreurs par sélection des métriques d'évaluation appropriées)

4.1.1 Construction de la référence

Le chapitre 3 nous a permis de mettre en évidence une méthode de production de référence qui permet de gagner en fiabilité dans les mesures cliniques qui peuvent être alors prises comme réalité terrain sous la forme d'intervalles de confiance. Du travail du chapitre précédent nous pouvons extraire deux types d'intervalles de confiance : un intervalle « expert » calculé depuis 3 modélisations indépendantes réalisées par 3 experts différents, et un intervalle « 3DS » correspondant à un intervalle plus étroit et donc plus fiable car calculé depuis 9 modélisations validées par 5 experts par le processus de l'entreprise partenaire. L'intervalle de confiance « 3DS » pour les mesures cliniques est notre réalité terrain pour l'évaluation de reconstructions 3D de colonne vertébrale, car il combine plusieurs profils d'experts à la fois en modélisations initiales et en modélisations validées.

4.1.2 Représentativité des données

La base de données utilisée pour évaluer est celle utilisée au chapitre 3, avec l'avantage qu'elle a été caractérisée de plusieurs manières : selon l'âge des patients, selon la sévérité

frontale des déformations, et selon la visibilité des images par vue ou par région. Cette caractérisation nous permet de pouvoir diviser notre échantillon de 64 patients en plusieurs sous échantillons afin de donner davantage d'informations pendant l'évaluation sur la répartition des erreurs parmi les sous échantillons.

4.2 Aspects cliniques de la reconstruction 3D de colonne vertébrale

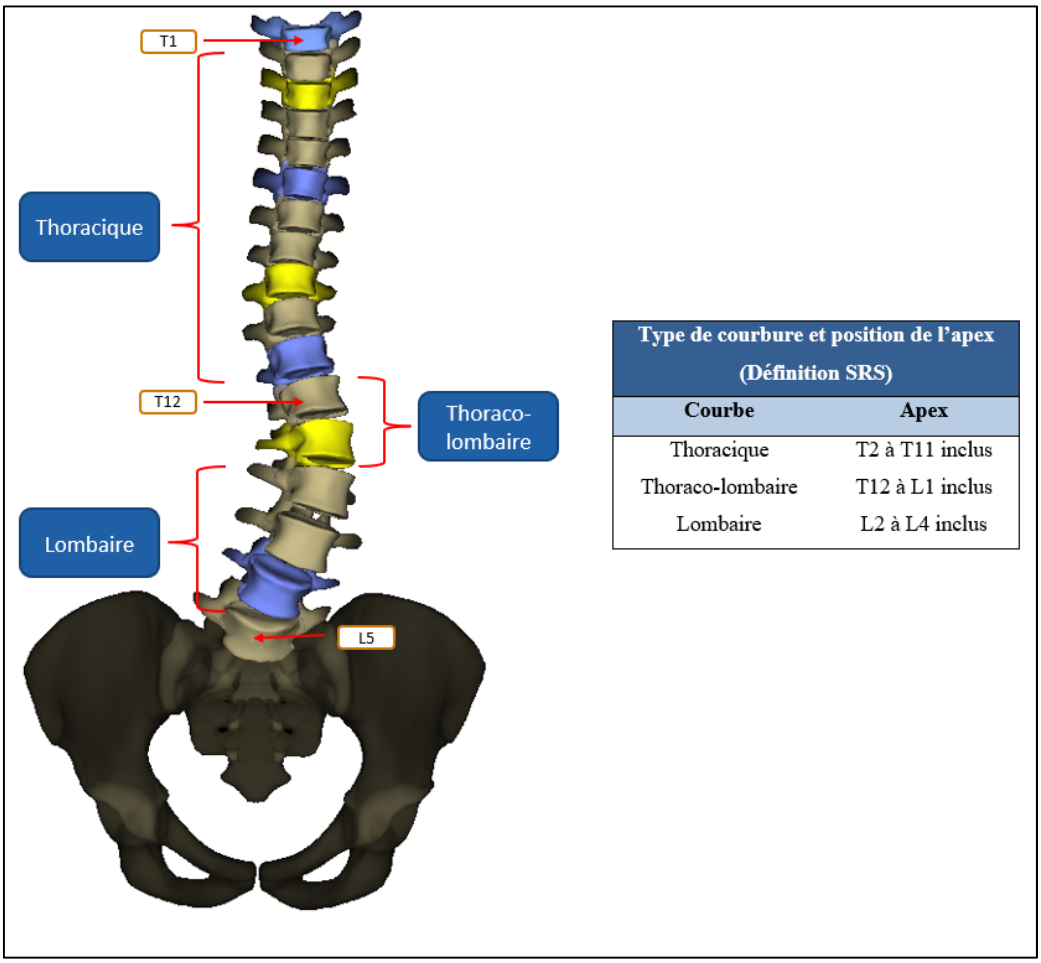
Nous nous sommes rapprochés des sociétés savantes pour identifier les aspects cliniques importants pour l'évaluation des déformations de la colonne vertébrale. Nous nous sommes ainsi appuyés sur plusieurs documents comme le Manuel de Mesures Radiographiques de (O'Brien & Spinal Deformity Study Group, 2004), les recommandations 2016 de la Société internationale d'Orthopédie, Scoliose et Traitement de Rééducation (Negrini *et al.*, 2018), et les publications de la Société de Recherche sur la Scoliose (Schwab *et al.*, 2012; Terran *et al.*, 2013). Nous avons ainsi pu identifier les mesures cliniques d'évaluation des déformations les plus utilisées, mais aussi les méthodes de classification des pathologies et adapter ces informations à notre évaluation.

4.2.1 Classifications de la scoliose

La scoliose a été longuement étudiée dans la littérature, avec une volonté des praticiens de trouver un moyen de mieux catégoriser les patients pour mieux personnaliser leur prise en charge. Ces classifications permettent d'orienter la prise en charge médicale des patients (du diagnostic jusqu'à la décision de chirurgie), ainsi les erreurs de la méthode à évaluer peuvent avoir un impact sur la classification du patient et donc sur sa prise en charge potentielle. Notre méthode d'évaluation propose de vérifier si les mesures réalisées par la méthode à évaluer classifie les patients de la même manière que les mesures composant notre référence. Cela permet par la suite de hiérarchiser les erreurs de reconstructions 3D selon leurs impacts sur ces classifications et sur leur gravité vis-à-vis du patient.

4.2.1.1 Les types de courbures selon la SRS

La SRS a défini une méthode de classification des courbures dans le plan frontal selon la position de la vertèbre sommet (l'apex) (O'Brien & Spinal Deformity Study Group, 2004) (voir figure 4.2).



Type de courbure et position de l'apex (Définition SRS)	
Courbe	**Apex**
Thoracique	T2 à T11 inclus
Thoraco-lombaire	T12 à L1 inclus
Lombaire	L2 à L4 inclus

Figure 4.2 Classification des courbures selon la position de l'apex (SRS), sur cet exemple, il existe deux courbures thoraciques et une courbure thoraco-lombaire (apex en jaune)

4.2.1.2 Les classifications pour les scolioses adolescentes

La classification la plus utilisée pour guider les traitements de déformations scoliotiques pour les adolescents est la classification de Lenke (Slattery & Verma, 2018). Cependant,

elle requiert plusieurs types d'images dont des radiographies où le patient est penché sur le côté, ce qui ne correspond pas aux données que l'on souhaite pouvoir évaluer, à savoir des radiographies bi-planaires avec le patient debout et droit et les modélisations 3D issues de ces images. Il existe cependant des recommandations par l'intermédiaire de la SOSORT[3], qui dans une revue très complète des pratiques en 2016 a pu établir des seuils cliniques correspondant à des objectifs de traitement en fonction de la sévérité des déformations frontales (voir tableau 4.1) (Negrini *et al.*, 2018).

Tableau 4.1 Objectifs de traitement selon la gravité des courbures (SOSORT)

SOSORT traitements pédiatriques visés	
Gravité de courbure	**But du traitement / Seuils**
Faible	Rester en dessous de 20° d'angle de Cobb maximal
Modérée	Rester en dessous de 30° d'angle de Cobb maximal
Sévère	Rester en dessous de 45° d'angle de Cobb maximal

4.2.1.3 Les classifications pour les adultes

La classification des déformations rachidiennes pour les adultes la plus utilisée est la classification SRS-Schwab (Terran *et al.*, 2013), qui permet de donner des informations supplémentaires sur l'équilibre sagittal du patient, c'est-à-dire l'inclinaison du tronc en avant ou en arrière, la morphologie du bassin, et la position du bassin par rapport à la colonne également (voir figure 4-3). Elle est composée de 4 critères qui sont le type de courbure (selon la position SRS de l'apex et le dépassement du seuil de 30 degrés d'angle de Cobb en vue frontale), la valeur de l'inclinaison pelvienne (PT), la valeur de l'axe vertical sagittal (SVA) et la valeur de l'harmonie lombo-pelvienne (PI-LL, la différence entre l'incidence pelvienne PI et la lordose lombaire L1/L5) qui permet d'évaluer la normalité de la position du bassin par rapport à la colonne lombaire (section 1.2.2.2).

[3] Société Scientifique Internationale sur les traitements orthopédiques et conservateurs de la Scoliose

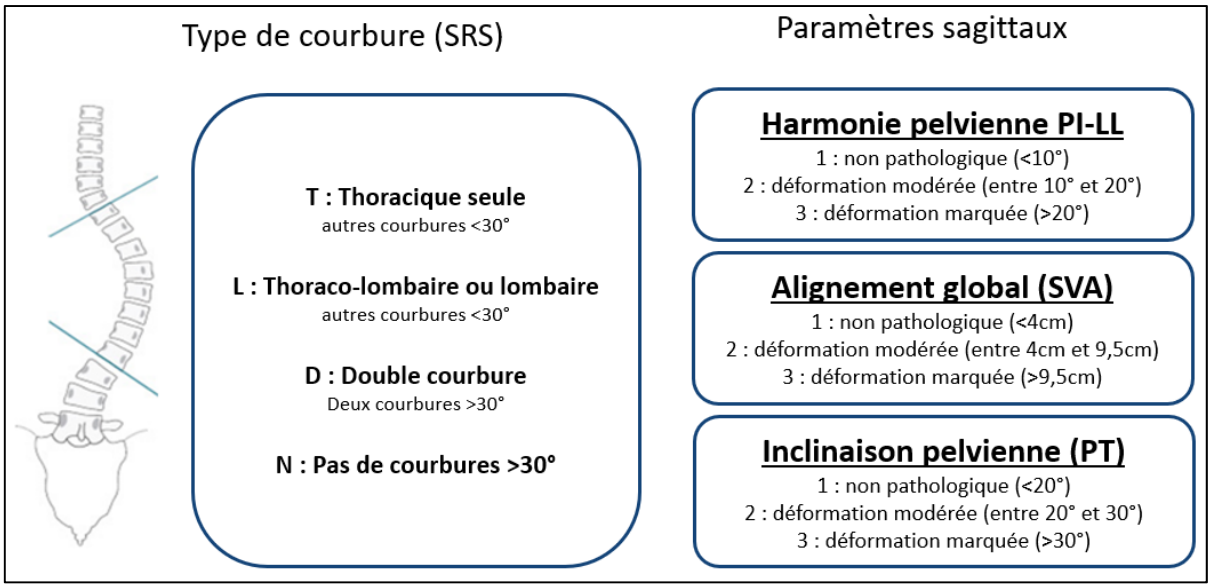


Figure 4.3 Classification SRS-Schwab pour les adultes
Adapté de Terran *et al.* (2013, p. 560)

4.2.1.4 Seuils de gravité pour l'angle de Cobb

L'angle de Cobb est une mesure importante pour évaluer la gravité des déformations scoliotiques. Nous nous inspirons des travaux de Negrini *et al.* (2018) sur les seuils de gravité basés sur l'angle de Cobb pour les adolescents, pour fixer 5 seuils :

- 10° : seuil de diagnostic
- 20° : seuil constituant la limite entre des déformations légères et des déformations modérées
- 30° : seuil constituant un objectif de traitement (le traitement consiste à conserver des courbures de moins de 30° d'angle de Cobb
- 40° : seuil constituant la limite entre des déformations modérées et des déformations sévères
- 45° : seuil de chirurgie

Nous proposons d'utiliser ces seuils pour créer une classification de gravité des déformations coronales pour toute la population y compris les adultes. Cela permet de renseigner la performance de mesure de l'angle de Cobb pour tous les types d'images.

4.2.2 Gravité des erreurs de mesures

En appliquant chaque classification à notre base de données d'évaluation d'une part et aux mesures issues de la méthode à évaluer d'autre part, nous pouvons analyser le classement donné par la méthode à évaluer, afin de caractériser le type d'erreur rencontré. La hiérarchie des erreurs a été décrite comme suit pour toutes les classifications, selon les scénarios d'erreurs décrits dans la figure 4.4 (dans cet exemple, seules 4 classes sont illustrées, avec pour les deux scénarios un résultat dans la classe 1, et un résultat dans la classe 3 respectivement). Ce sont des règles générales que l'on peut appliquer par la suite aux classifications et seuils cliniques.

- Une erreur MINEURE décrit une erreur de mesure n'entrainant pas d'erreur de classification (la mesure n'est pas dans l'intervalle de confiance de mesure mais n'entraine pas de changement de classe);
- Une erreur MAJEURE décrit une erreur de mesure entrainant une erreur de classification sans impliquer de chirurgie et avec seulement une classe d'écart avec

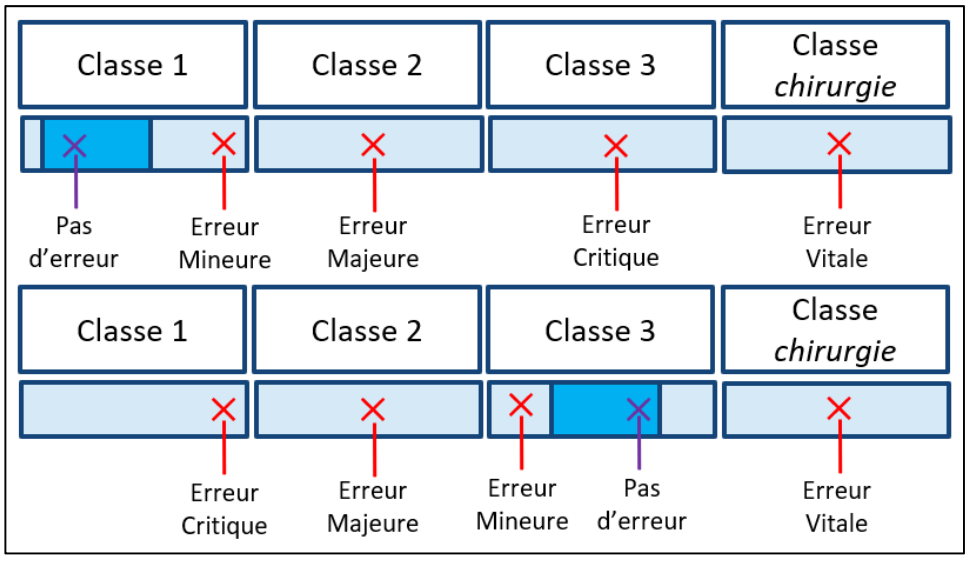


Figure 4.4 Scénarios d'erreurs de classifications ; l'encadré bleu foncé correspond aux intervalles de confiances de mesures cliniques pris pour référence et utilisés pour la classification, les erreurs de mesures peuvent être hiérarchisées selon le classement donné par la mesure de la méthode évaluée (Croix rouge pour une mesure hors de l'intervalle de confiance de référence, croix mauve pour une mesure dans l'intervalle de confiance de référence)

la référence (la mesure n'est pas dans l'intervalle de confiance de mesure et entraine un changement de classe);

- Une erreur CRITIQUE décrit une erreur de mesure entrainant une erreur de classification sans impliquer de chirurgie et avec au moins deux classes d'écart avec la référence (la mesure n'est pas dans l'intervalle de confiance de mesure et entraine deux changements de classe sans atteindre la classe de chirurgie);
- Une erreur VITALE décrit une erreur de mesure entrainant une erreur de classification impliquant de la chirurgie, quel que soit le nombre de classes d'écart avec la référence (la mesure n'est pas dans l'intervalle de confiance de mesure et atteint la classe de chirurgie alors qu'elle ne devrait pas d'après la référence ou ne l'atteint pas alors qu'elle devrait d'après la référence);

4.2.2.1 Gravité des erreurs pour les adolescents

Depuis les objectifs de traitement décrits par Negrini *et al.* (2018) (section 4.2.1.2), qui ciblent l'angle de Cobb maximal observé, nous avons adapté les erreurs de classifications décrites à la section 4.2.1.2 (voir tableau 4.2). Cette table d'impact formalise les types d'erreurs (mineur, majeur, critique et vital) précédemment définis. Si la méthode évaluée ne positionne pas la mesure de l'angle de Cobb dans l'intervalle de confiance de référence mais se classe dans le bon échelon de mesure de Cobb, l'erreur sera toujours mineure. Si l'erreur de position en dehors de l'intervalle de confiance de référence implique un changement de classe incluant le seuil de 45°, elle sera toujours vitale car c'est la classe de chirurgie. Enfin, pour les trois autres classes, les erreurs sont définies selon le nombre de classe d'écart entre celle donnée par la valeur de l'angle de Cobb de la méthode évaluée et celle donnée par l'intervalle de confiance de la référence.

4.2.2.2 Gravité des erreurs pour les adultes

Pour la classification SRS-Schwab, la littérature ne nous permet pas de formaliser des erreurs vitales aussi clairement que pour les adolescents, les impacts cliniques sont donc

restreints aux impacts mineurs, majeurs et critiques, pour le type de courbure d'un côté et chaque paramètre sagittal de l'autre (voir tableau 4.3 et section 4.2.1.3).

Tableau 4.2 Table d'impact pour la mesure du plus grand angle de Cobb selon les critères de la SOSORT, et les échelons de mesures définis par les seuils de référence

Impact clinique pour une erreur sur le plus grand angle de Cobb		Seuil de référence			
		>45°	30°-45°	20°-30°	<20°
Seuil évalué	>45°	Mineur	Vital	Vital	Vital
	30°-45°	Vital	Mineur	Majeur	Critique
	20°-30°	Vital	Majeur	Mineur	Majeur
	<20°	Vital	Critique	Majeur	Mineur

Tableau 4.3 Types d'impacts pour la classification SRS-Schwab pour les adultes

Impact clinique des erreurs sur le type de courbure		Classification SRS-Schwab de la référence			
		N	T	L	D
Classification SRS-Schwab de la méthode évaluée	N		Majeur	Majeur	Critique
	T	Majeur		Mineur	Majeur
	L	Majeur	Mineur		Majeur
	D	Critique	Majeur	Majeur	

Impact clinique des erreurs sur PT/SVA/PI-LL		Classification SRS-Schwab de la référence		
		1	2	3
Classification SRS-Schwab de la méthode évaluée	1	Mineur	Majeur	Critique
	2	Majeur	Mineur	Majeur
	3	Critique	Majeur	Mineur

Pour le type de courbure, 4 classes sont définies, avec les courbures N (aucune courbure au-dessus de 30°), la classe T (une seule courbure thoracique au-dessus de 30°), la classe L (une seule courbure lombaire ou thoraco-lombaire au-dessus de 30°), et la classe D (deux courbures au-dessus de 30°). Le type de courbure n'étant pas un intervalle de confiance de mesure, il ne peut y avoir d'erreurs mineures sur le même principe que la gravité des erreurs pour les adolescents et l'angle de Cobb. Néanmoins, nous proposons de les définir comme la mauvaise identification en cas de courbure unique (soit une courbure Thoracique quand elle est définie Lombaire dans la référence et inversement). Une erreur majeure consiste à ne pas identifier une courbure unique lorsqu'il y en a une, ou à ne pas identifier une deuxième

courbure lorsqu'il y en a deux. Une erreur critique serait de ne pas identifier de courbures alors qu'il y en a deux, ou d'identifier deux courbures alors qu'il n'y en a aucune.

Pour les paramètres sagittaux, 3 classes sont définies selon la valeur de chaque mesure : 1 correspond à une déformation non pathologique, 2 à une déformation modérée, et 3 à une déformation sévère. Sur les mêmes définitions présentées à la section (4.2.2), une erreur mineure, majeure ou critique correspondra respectivement à une mesure hors de l'intervalle de confiance de référence mais sans écart de classe, avec un seul écart de classe, ou deux écarts de classe.

La gravité d'erreur finalement retenue pour un sujet sera la plus haute atteinte parmi tous les paramètres de classification.

4.2.2.3 Gravité des erreurs sur l'angle de Cobb

Des différents seuils décrits pour l'angle de Cobb (section 4.2.1.4), nous pouvons de nouveau formaliser une table d'impacts selon le classement donné par la mesure de la méthode à évaluer par rapport au classement donné par la valeur de référence (voir tableau 4.4). Nous proposons une hiérarchie des erreurs ajustée par rapport aux définitions de la figure 4.4, afin de mieux refléter les pratiques cliniques. Ainsi, une erreur entrainant un écart d'une classe est mineure, à l'exception des erreurs de diagnostic qui seront notées majeures (quand la méthode évaluée ne détecte pas de courbures au-dessus de 10° alors qu'il y en a et vice versa). Une erreur majeure correspond à une erreur de mesure entrainant un changement de deux classes, à l'exception des erreurs de traitement qui seront notées critiques (quand la méthode évaluée donne une valeur de Cobb au-dessus de 30° alors que la valeur de référence n'atteint pas 30° et vice versa). Une erreur critique correspond à une erreur entrainant un changement de plus de deux classes sans impliquer la classe de chirurgie ou impliquant une classe de traitement (30° ou 40°). Une erreur vitale correspond à une erreur impliquant la classe de chirurgie (>45°).

Tableau 4.4 Table d'impacts pour les seuils de gravité de l'angle de Cobb

Impacts selon valeurs de l'angle de Cobb		Référence					
		<10°	>10°	>20°	>30°	>40°	>45°
Méthode évaluée	<10°		Majeur	Majeur	Critique	Critique	Vital
	>10°	Majeur		Mineur	Critique	Critique	Vital
	>20°	Majeur	Mineur		Majeur	Critique	Vital
	>30°	Critique	Critique	Majeur		Mineur	Vital
	>40°	Critique	Critique	Critique	Mineur		Vital
	>45°	Vital	Vital	Vital	Vital	Vital	

4.3 Méthode d'évaluation

Maintenant que nous avons mis en place notre méthode pour inclure les aspects cliniques à notre méthode d'évaluation (choix et adaptations des classifications cliniques, création des tables d'impacts), nous allons pouvoir exposer les différentes composantes de la méthode d'évaluation.

4.3.1 Mesurer la performance

Nous utilisons deux types de mesures de performance. Pour le premier, nous nous appuyons sur notre revue de littérature pour identifier la performance de mesures cliniques de la méthode de reconstruction 3D à évaluer, par rapport à des intervalles de confiance qui constituent notre réalité terrain (l'intervalle « 3DS » défini à la section 4.1.1). Pour le second, nous proposons une nouvelle manière de mesurer la performance en nous appuyant sur la littérature clinique et sur les classifications de pathologies.

4.3.1.1 Performance de mesures cliniques

Pour quantifier les erreurs de mesures cliniques nous allons utiliser le taux de succès, le MAD, le SMAPE (définitions en section 1.4.4), et les mesures de précision comme la sensibilité et la spécificité (définitions en section 1.4.1). Ces métriques présentent toutes l'avantage de pouvoir être calculées pour l'échantillon complet ou pour les sous-catégories de sévérité, de visibilité et d'âge, et sont calculés pour l'angle de Cobb maximal ainsi que

pour les paramètres de balance sagittale (T1/T12, T4/T12, L1/L5 et L1/S1) et les paramètres pelviens (PT, PI et SS).

4.3.1.2 Performances cliniques

Notre méthode d'évaluation inclut les aspects cliniques de la reconstruction 3D de colonne vertébrale en s'appuyant sur des classifications cliniques de gravité des erreurs. Cela n'a pas pour but de pouvoir donner une valeur diagnostique à la méthode évaluée, mais simplement de hiérarchiser les erreurs d'un point de vue clinique, par l'intermédiaire des tables d'impacts.

Pour ce faire, il a fallu caractériser notre base de données d'évaluation en attribuant à chaque patient la classe qui lui est associée selon chacune des classifications définies aux sections 4.2.1 et 4.2.2. Certains patients ont toutefois été problématiques. En effet, il est possible d'observer des chevauchements des intervalles de confiance sur deux classes différentes (voir figure 4.5). Pour pallier ce problème, nous mettons en place une nouvelle métrique d'évaluation qui mesure l'accord expert (AE) entre la méthode à évaluer et la réalité terrain (équation 4.1). Lorsque dans la référence il n'y a pas de chevauchements de classes, il sera possible d'atteindre 100% d'accord avec la méthode que l'on évalue. Si dans la réalité terrain il y a 3 mesures validées sur 9 qui donnent la classe 1 et les 6 autres qui

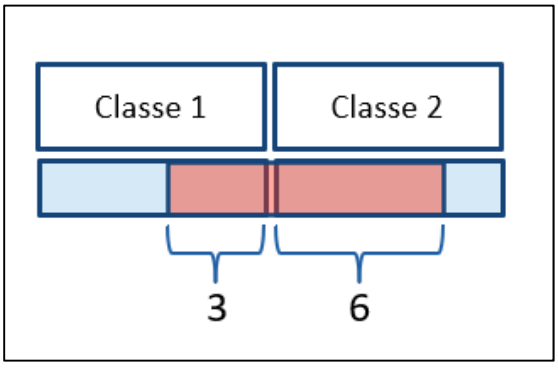


Figure 4.5 Exemple de chevauchement de classe par l'intervalle de confiance de la réalité terrain (en rouge), avec le nombre associé de modélisations validées

donnent la classe 2 pour un patient donné, la méthode évaluée obtiendra 33% d'accord expert pour une mesure clinique donnant la classe 1 (voir figure 4.5).

$$AE = \frac{\sum_{m}^{n} \frac{m_{accord}}{9} \times 100}{n} \tag{4.1}$$

Avec m_{accord} le nombre de modélisations validées en accord avec la mesure évaluée

Un nouveau sous échantillon de sujets est également créé afin de séparer les sujets pour lesquels il n'y a aucune divergence dans la référence et les sujets pour lesquels il existe des divergences. Ce sous échantillon permet pour une évaluation de communiquer une information très pertinente. On peut en effet savoir, quand il y a des erreurs avec impacts cliniques, si ces erreurs sont réparties dans le sous échantillon avec une référence sûre et fiable à 100% ou bien si ces erreurs font partie du sous échantillon de référence où il existe des désaccords entre experts, ce qui limiterait leur impact sur l'évaluation finale.

4.3.2 Expérimentations de la méthode d'évaluation

Nous avons réalisé des évaluations pour deux méthodes de reconstruction 3D, soit l'approche de Humbert *et al.* (2009), qui est celle utilisée pour réaliser l'étude du chapitre 3 sur la construction de la référence, et celle de Aubert *et al.* (2019). La première est une méthode semi-automatique de reconstructions 3D de la colonne vertébrale, qui requiert la participation d'un opérateur. Nous disposons pour cela de 64 modélisations réalisées séparément par deux experts isolés, SR2 et JR3 (elles correspondent aux modélisations réalisées par ces deux experts à l'étape d'analyse du chapitre précédent). La seconde est une méthode automatique de reconstructions 3D de colonne vertébrale. Nous disposons pour cela de 64 modélisations issues d'un algorithme automatique de reconstructions 3D, que nous appelons NFCU. Le but ici n'est pas d'évaluer NFCU, mais de montrer l'intérêt de notre méthode d'évaluation et ce qu'elle apporte de nouveau dans des évaluations de performances.

- **Taux de succès**

Le tableau 4.5 montre le taux de succès pour JR3, SR2 et NFCU sur les mesures cliniques issues des reconstructions 3D. SR2 obtient de meilleurs taux de succès que JR3. L'algorithme NFCU obtient de moins bons taux de succès, avec son moins bon taux de 47% pour T1/T12 et son meilleur taux pour la pente sacrée SS (80%).

Tableau 4.5 Taux de succès des mesures cliniques pour les experts JR3 et SR2, et pour NFCU

	IC	Cobb Max	T1/T12	T4/T12	L1/L5	L1/S1	PI	SS	PT
JR3	Réalité terrain	92%	91%	91%	92%	84%	97%	95%	94%
SR2	Réalité terrain	97%	97%	98%	97%	95%	95%	94%	100%
NFCU	Réalité terrain	73%	47%	67%	70%	70%	80%	78%	77%

- **MAD**

La figure 4.6 montre la variation du MAD pour JR3, SR2 et de l'algorithme automatique NFCU, pour les mesures cliniques issues des reconstructions 3D. Ce type de représentation nous permet d'analyser l'amplitude des écarts à la référence (ici le centre des intervalles de confiances constituant notre réalité terrain, obtenu par le calcul de la moyenne des mesures validées) : pour chaque mesure clinique, plus la courbe est proche du centre, plus les mesures sont proches du centre des intervalles de confiance de référence. JR3 a tendance à légèrement surévaluer les mesures cliniques par rapport à SR2, avec des MAD légèrement plus grandes (plus éloignées du centre) (figure 4.6a et 4-6c). SR2 semble avoir les mesures cliniques les plus proches des centres des intervalles de confiance, globalement pour tous les patients, adultes ou adolescents, sévères ou modérés (figure 4.6c et 4-6d), avec des MD systématiquement inférieurs à 2°. L'observation des données pour la pédiatrie et pour les adultes nous permet de voir également que les plus grandes erreurs sont faites dans la population des adultes, surtout pour l'algorithme NFCU. Le MAD ne permet toutefois pas de savoir si la mesure évaluée se situe dans l'intervalle de confiance de référence, et permet seulement de quantifier les écarts au centre de l'intervalle de confiance.

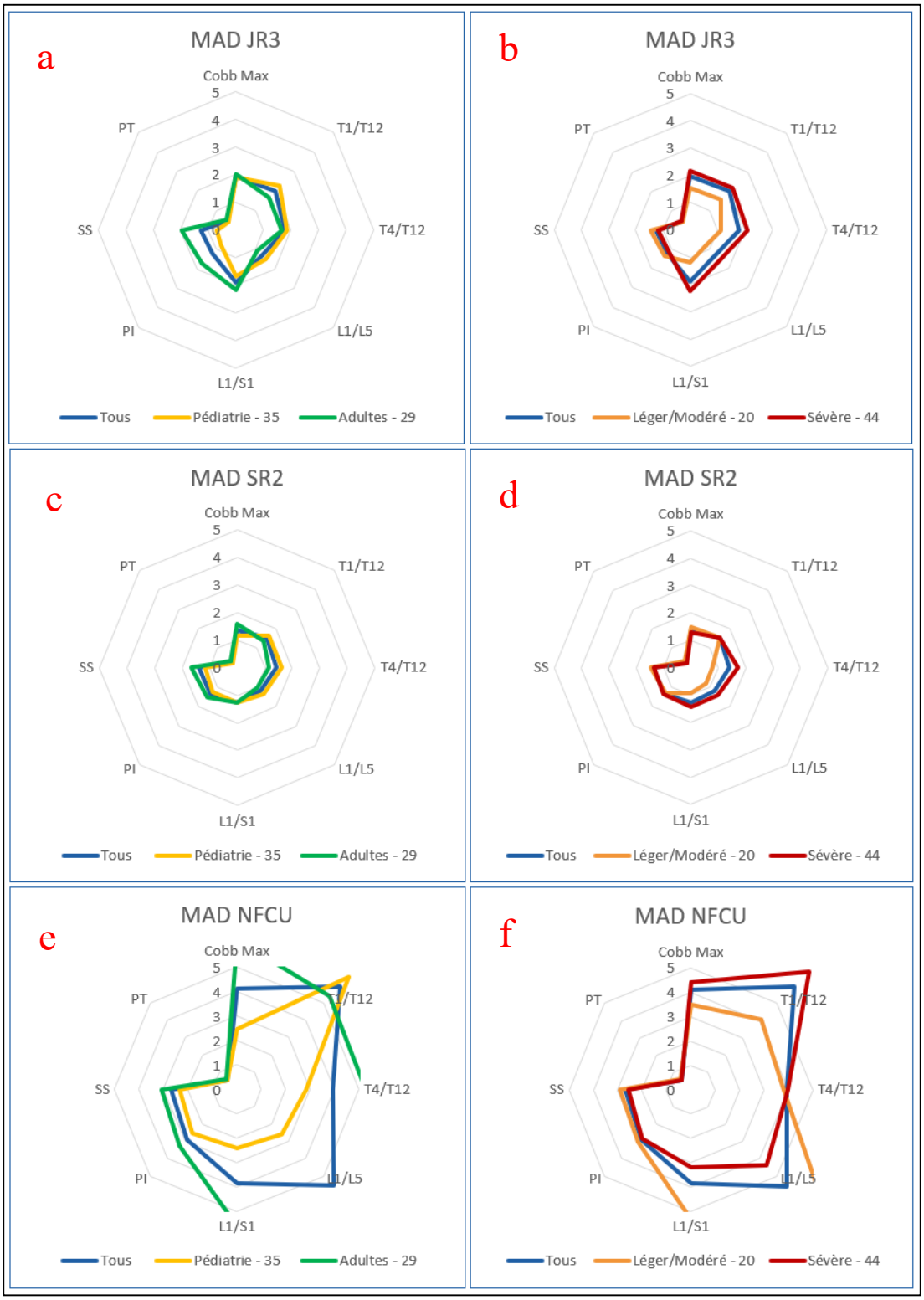


Figure 4.6 Moyenne absolue des différences (MAD) pour les reconstructions 3D des experts JR3 (a ; b) et SR2 (c ; d), et pour l'algorithme automatique de reconstruction 3D NFCU (e ; f) selon deux catégorisations de population différentes, l'âge (à gauche) et la sévérité (à droite)

- **SMAPE des experts JR3, SR2 et de la méthode automatique NFCU**

Le tableau 4.6 montre les résultats pour la performance SMAPE des experts JR3 et SR2, et de l'algorithme automatique NFCU, pour quelques catégories de patients et de paramètres cliniques. Décrit à la section 1.4.4, le SMAPE permet d'obtenir un pourcentage d'erreur pour plusieurs mesures cliniques combinées ou pour une seule. Les valeurs de référence prises ici sont les centres des intervalles de confiance pour chaque mesure clinique. Plus ce pourcentage est haut, plus il existe d'erreurs de mesures.

On peut voir que les deux experts ont des performances très similaires, avec plus de difficulté pour l'évaluation de l'angle de Cobb maximal des patients avec déformations modérées que pour les autres paramètres cliniques (9% pour JR3, 11% pour SR2 et 16% pour NFCU de performance SMAPE pour le Cobb maximal). Si l'on regarde les résultats pour l'algorithme NFCU, les performances sont moins bonnes, surtout pour les adultes (7 à 29% d'erreur SMAPE contre 2 à 7% pour les adolescents).

Tableau 4.6 Performance SMAPE pour les reconstructions 3D des experts JR3 et SR2 et pour l'algorithme automatique de reconstruction 3D NFCU, pour la mesure du Cobb maximal, pour les paramètres pelviens (PP) et pour la balance sagittale (BS)

Méthode évaluée	JR3			NFCU			SR2		
Catégorie de sujets	Cobb max	PP	BS	Cobb max	PP	BS	Cobb max	PP	BS
Tous	4%	3%	5%	8%	5%	17%	4%	4%	4%
Pédiatrie	2%	2%	4%	2%	4%	7%	1%	2%	4%
Adultes	7%	4%	6%	16%	7%	29%	8%	6%	5%
Léger/Modéré	9%	3%	3%	16%	7%	30%	11%	4%	2%
Sévère	2%	3%	6%	5%	5%	11%	1%	3%	5%

- **Performances de classification**

Le tableau 4.7 montre les résultats obtenus par JR3, SR2, et l'algorithme NFCU sur l'identification de gravité du plus grand angle de Cobb identifié, avec les accords experts atteints en pourcentage (voir section 4.3.1.2, équation 4.1) et le nombre d'impacts critiques

et vitaux (voir section 4.2.2). L'algorithme NFCU présente les plus mauvais résultats, avec 5 erreurs critiques et vitales et un accord expert de 82% contre 3 et 2 erreurs vitales pour JR3 et SR2 respectivement ainsi que 86% et 90% d'accord expert. À la différence des experts, NFCU obtient une majorité d'erreurs pour la population adulte, ce qui se reflète sur son pourcentage d'accord expert (67% d'accord pour les adultes contre 90% pour les adolescents). Globalement, nous observons de plus mauvais résultats pour les adultes et pour la population avec des déformations légères à modérées.

Tableau 4.7 Accord expert atteint et nombres d'impacts critiques et vitaux pour les reconstructions 3D des experts JR3 et SR2 et pour l'algorithme automatique de reconstruction 3D NFCU selon les seuils de gravité du plus grand angle de Cobb

Seuils de gravité pour l'angle de Cobb	**JR3**			**NFCU**			**SR2**		
Catégorie - n	Accord expert	Impacts Critiques	Impacts Vitaux	Accord expert	Impacts Critiques	Impacts Vitaux	Accord expert	Impacts Critiques	Impacts Vitaux
Tous - 64	86%	0	3	82%	3	2	90%	0	2
Pédiatrie - 35	90%	0	2	90%	0	1	90%	0	2
Adultes - 29	79%	0	1	67%	3	1	85%	0	0
Léger/Modéré - 20	80%	0	0	68%	1	0	83%	0	0
Sévère - 44	89%	0	3	88%	2	2	93%	0	2

Le tableau 4.8 montre les accord experts obtenus pour la classification des adolescents issue des recommandations de la SOSORT, pour la totalité du sous échantillon de la pédiatrie (35 sujets) ainsi que pour le sous échantillon sans chevauchement de classes dans la réalité terrain (30 sujets) et le sous échantillon avec chevauchement de classes (5 sujets). SR2 fait un moins bon score d'accord experts pour les cas avec chevauchement de classes (73%) que JR3 (97%) tandis que NFCU obtient un accord expert de 90%.

Le tableau 4.9 montre les résultats d'évaluation pour la performance de classification des adultes pour les reconstructions 3D de JR3, SR2 et l'algorithme NFCU, en particulier l'accord expert atteint ainsi que le nombre d'erreurs mineures, majeures et critiques identifiées. L'algorithme NFCU présente le moins bon pourcentage d'accord expert (47%) ainsi que le plus grand nombre d'erreurs majeures et critiques (3 erreurs majeures et 2

erreurs critiques). Les reconstructions 3D des deux experts présentent presque les mêmes nombres d'erreurs, les mesures cliniques de JR3 ayant entrainé une erreur critique de plus que SR2.

Tableau 4.8 Accord expert atteint pour les reconstructions 3D des experts JR3 et SR2, de l'algorithme automatique NFCU, pour la classification des adolescents, selon le chevauchement de classes dans la réalité terrain

Accord expert pour la classification des adolescents	Tous	Sujets sans chevauchement de classes	Sujets avec chevauchement de classes
n	35	30	5
JR3	94%	97%	87%
SR2	93%	97%	73%
NFCU	94%	97%	90%

Tableau 4.9 Accord expert et type d'erreurs de mesures pour l'évaluation de la performance de classification des adultes des reconstructions 3D de JR3, SR2 et l'algorithme NFCU

	Accord expert pour la classification des adultes	Nombre d'erreurs mineures	Nombre d'erreurs majeures	Nombre d'erreurs critiques
SR2	55%	14	2	0
JR3	55%	14	2	1
NFCU	47%	14	3	2

4.3.3 Discussion

Grâce à notre méthode d'évaluation, nous sommes capables de quantifier des erreurs de mesures cliniques, de les qualifier selon leur gravité d'un point de vue clinique, et cela par rapport à une référence considérée comme réalité terrain.

Nous avons évalué la performance des reconstructions 3D de deux experts ayant travaillé avec l'approche de Humbert *et al.* (2009). Ces deux experts (SR2 qui est un expert senior avec plus d'expérience de modélisations 3D que l'expert junior JR3), ont tous deux participé au processus de reconstructions 3D mis en place au chapitre 3 pour construire la réalité terrain. Cela constitue un biais d'évaluation puisque nous avons évalué leurs reconstructions 3D déjà utilisées pour construire la référence (elles correspondent aux

reconstructions 3D issues de l'étape 1 de reconstruction 3D tel que décrit à la section 3.2.1.1). Malgré ce biais, nous pouvons mettre en évidence des différences entre deux experts isolés, avec JR3 par exemple qui a des taux de succès moins bons que SR2 pour les mesures classiques (tableau 4.5). Ces différences n'occultent cependant pas le fait que l'expert junior obtient des mesures cliniques issues de ses reconstructions 3D avec des taux de succès supérieurs à 84%. On peut voir également dans le tableau 4.8 de performance SMAPE qu'il semble y avoir plus de difficulté pour des experts d'identifier le plus grand angle de Cobb chez les adultes et chez les sujets avec déformations légères ou modérées. Ces deux catégories sont liées car l'échantillon de pédiatrie inclut une grande majorité de sujets sévères (Cobb moyen de 58.5° ± 13.3°) alors que l'échantillon des adultes inclut une majorité de sujets avec déformations légères ou modérées (Cobb moyen de 31.5 degrés ± 19°). Pour la mesure de l'angle de Cobb, on pourrait donc voir une limite de l'approche de Humbert *et al.* (2009) dans la correcte identification des sévérités frontales de la scoliose pour les adultes. Or, la littérature clinique ne place pas l'angle de Cobb au centre de la prise en charge des patients comme c'est le cas pour les adolescents. Les investigations de scolioses chez les adultes s'intéressent de manière égale à des paramètres autres que l'angle de Cobb, ce qui veut dire que l'évaluation de performance d'un point de vue clinique pour les méthodes de reconstructions 3D d'images issues de population adulte doit être ajustée dans l'interprétation aux caractéristiques de cette population.

Pour l'évaluation de l'algorithme NFCU résultant de l'approche de Aubert *et al.* (2019), nous avons vu que la performance ne dépasse pas 80% de taux de succès pour les mesures cliniques standards. Les écarts au milieu des intervalles de confiance de notre réalité terrain sont d'ailleurs très grands pour NFCU (figure 4.5), avec de nettes différences entre l'échantillon des adultes et l'échantillon des adolescents. Cela peut s'expliquer par le fait que la méthode automatique de Aubert *et al.* (2019) a été entrainée sur des sujets adolescents et non sur des adultes. Ce défaut se ressent dans le reste de l'évaluation (avec globalement de très bons résultats pour les adolescents pour le SMAPE (tableau 4.6), pour l'identification des classes de gravité selon l'angle de Cobb (tableau 4.7), et pour la classification SOSORT (tableau 4.8)), et de moins bons résultats pour l'échantillon adulte (tableau 4.9). La mise en parallèle des résultats d'évaluation de NFCU et des deux experts

JR3 et SR2 permet de positionner la performance de NFCU par rapport à des experts isolés, avec la catégorisation des sujets qui permet aux développeurs de mieux comprendre le comportement de NFCU et donc de proposer des améliorations appropriées.

4.4 Résumé

La nouvelle méthode d'évaluation que nous proposons présente les spécificités suivantes :

- des intervalles de confiances de mesures cliniques ainsi que leur centre (moyenne des mesures de référence), que nous pouvons qualifier de réalité terrain selon la définition de Cardoso *et al.* (2014). Nous avons en effet pu démontrer le niveau de confiance de cette référence dans notre travail du chapitre 3 ;
- une base de données d'évaluations dont nous avons décrit la représentativité selon l'âge des patients, la sévérité des déformations, et la visibilité des structures dans les images ;
- des mesures de performance pour quantifier les erreurs de mesures cliniques (Taux de succès, pourcentage d'erreur SMAPE, différence moyenne MAD, pourcentage d'accord expert) ;
- des mesures de performance pour qualifier les erreurs de mesures cliniques (caractérisation des types d'erreurs selon des tables d'impacts cliniques, basées sur la littérature et des classifications de pathologies officielles).

Le travail mené au chapitre 4 a permis de mettre en place une nouvelle méthode d'évaluation des reconstructions 3D de colonne vertébrale qui répond aux besoins de l'entreprise partenaire et de tester cette dernière pour l'évaluation de deux méthodes de reconstruction 3D. En effet, il n'existe pas d'étalon or pour évaluer la qualité des modélisations 3D, et jusqu'ici il n'y avait pas de recommandations quant à la mobilisation d'experts pour construire une référence. Le fait d'avoir pu combiner le travail du chapitre 3 et les aspects cliniques synthétisés au chapitre 4 permet d'avoir une évaluation à la fois fiable et pertinente. Il est possible non seulement d'identifier et de quantifier des erreurs de mesures cliniques issues de modélisations 3D mais aussi de les qualifier sur le type

d'images desquelles elles proviennent et sur leur hiérarchisation selon leur gravité clinique. La construction de cette méthode d'évaluation correspond à deux besoins identifiés dans la littérature et sur le terrain, à savoir le besoin d'avoir une référence de haute qualité et le besoin de redonner un but aux évaluations : peu importe d'avoir une méthode de segmentation performante dans 95% des cas si les 5% d'erreurs arrivent sur des cas sévères avec le plus haut niveau de gravité de ces erreurs. Pouvoir apporter ces informations dans un résultat d'évaluation entre plusieurs méthodes permettraient ainsi de mieux les hiérarchiser dans leur performance technique en incluant une performance clinique, qui ne peut se concevoir sans le concours de sociétés savantes ou de recommandations terrains officielles.

Pour notre domaine d'application spécifique, il manquerait cependant un aspect important dans l'évaluation de la performance des méthodes de modélisation 3D de la colonne vertébrale, ce sont les modèles 3D. En effet, le but initial de la construction de notre outil d'évaluation est de contribuer à améliorer les méthodes de reconstruction 3D. Aujourd'hui, les modèles 3D sont reconstruits avec une méthode semi-automatique (Humbert *et al.*, 2009) par un groupe d'experts selon le processus de l'entreprise partenaire décrit au chapitre 3. Les experts sont formés spécifiquement pour créer des modèles 3D reproductibles permettant d'accéder à des mesures cliniques. Il serait dans ce contexte intéressant d'analyser la reproductibilité des modèles 3D eux-mêmes, les comparer à la reproductibilité des mesures cliniques, et ouvrir ainsi des perspectives sur la construction de modèles 3D de référence que l'on pourrait inclure à notre méthode d'évaluation.

CHAPITRE 5

EXTENSION DE LA MÉTHODE DE CRÉATION DE RÉFÉRENCE AUX MODÈLES 3D DE COLONNE VERTÉBRALE

5.1 Introduction

Le travail du chapitre 3 nous a permis de voir qu'il est possible d'obtenir une référence définie par des mesures cliniques moyennes et leurs intervalles de confiance associés, issues de reconstructions 3D. Cette référence est plus fiable car les intervalles de confiance autour des mesures sont plus étroits, grâce à l'inclusion de contre vérifications et discussions entre experts. Cette référence n'est cependant en lien qu'avec les mesures cliniques extraites automatiquement des modèles 3D surfaciques et non directement avec les modèles 3D. Le travail du chapitre 4 nous a permis de mettre en place une méthode d'évaluation regroupant des mesures de performance sur les mesures cliniques en tenant compte des informations supplémentaires sur la gravité clinique des erreurs de performance.

Nous souhaitons dans ce chapitre ajouter une évaluation des modèles 3D eux-mêmes, et ne plus être restreint seulement aux mesures cliniques issues des modèles 3D. Le but de ce chapitre est donc d'analyser sur le même principe que le chapitre 3 la reproductibilité des modèles 3D et de les comparer à la reproductibilité des mesures cliniques. Cette analyse va nous permettre de reconnaître les éléments de qualité des modèles 3D qui ont une influence sur les paramètres cliniques, et ainsi donner des perspectives sur l'inclusion de ces résultats à notre méthode d'évaluation développée au chapitre 4, en utilisant les modèles 3D pour la création de référence.

5.2 Méthodes

5.2.1 Données

Sur le même échantillon de patients qu'au chapitre 3, nous récupérons les positions de plusieurs points d'intérêt des modèles 3D depuis le logiciel de recherche du laboratoire

(SterLIO©) : pour chaque vertèbre, le centre du corps vertébral, le centre de chaque pédicule, l'extrémité du processus épineux, le centre et le coin des plateaux (voir figure 5.1). Nous récupérons également les orientations du corps vertébral et des plateaux de chaque vertèbre. Ces orientations sont estimées par rapport au référentiel de chaque vertèbre suivant les définitions de (Stokes, 1994).

Nous réalisons cette opération pour chaque étape de modélisation : l'étape de reconstruction 3D, l'étape de revue et l'étape de validation (voir figure 1.11).

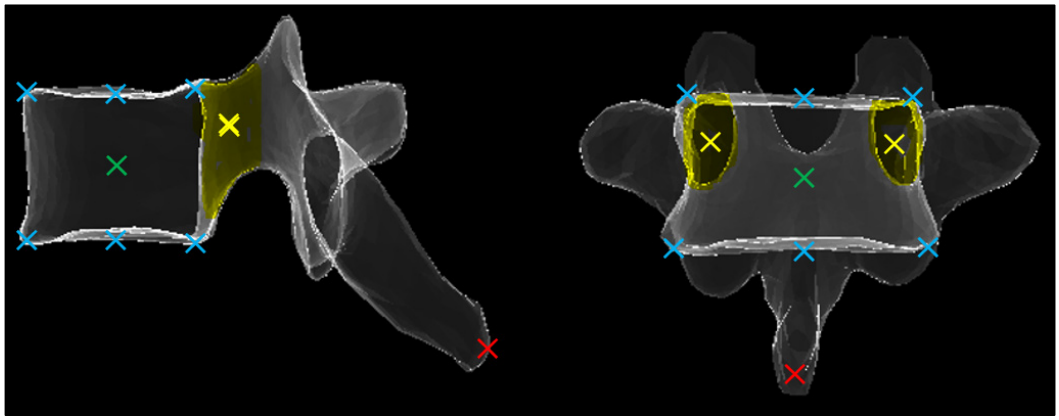


Figure 5.1 Points d'intérêts de chaque vertèbre : centre du corps vertébral (croix vertes), centre des pédicules (croix jaunes), processus épineux (croix rouges) et points des plateaux (croix bleues)

Les coordonnées 3D représentent la position des points en X, Y et en Z, et les orientations L, S et A correspondent à l'inclinaison des structures dans le plan frontal, dans le plan sagittal ou dans le plan axial (voir figure 5.2) (Skalli *et al.*, 1995).

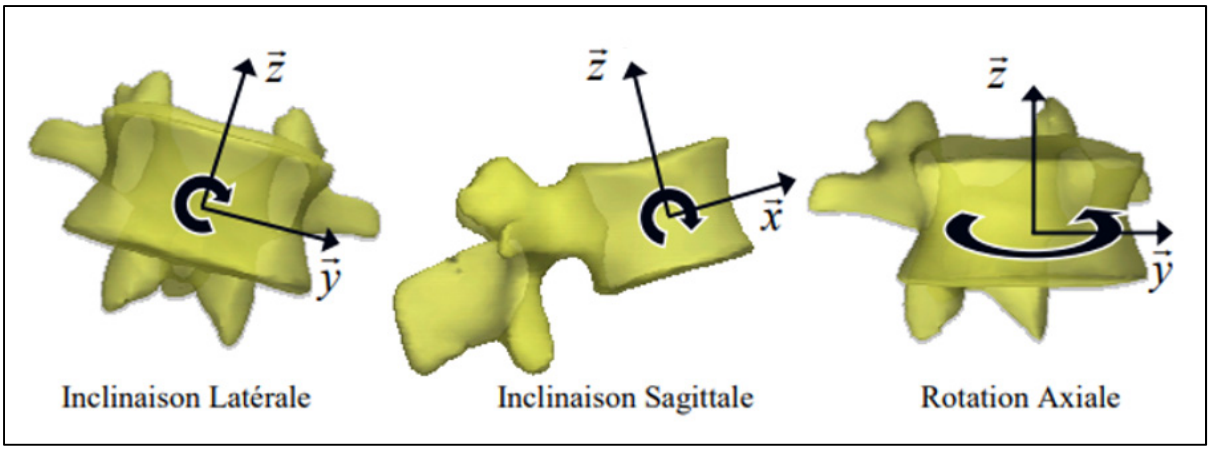


Figure 5.2 Référentiel vertébral, position et orientation dans les différents plans anatomiques
Adapté de Skalli *et al.* (1995)

5.2.2 Analyses statistiques

L'estimation de la précision de la position des points va permettre de montrer si la méthode de production de référence permet d'obtenir aussi bien une réalité terrain pour les mesures cliniques qu'une réalité terrain pour les modèles 3D. En effet, si l'ensemble des points du modèle 3D démontrent la même augmentation de précision avec la méthode de production de référence, il sera possible d'utiliser la position des points comme référence d'évaluation. En revanche, si certains points du modèle ne démontrent pas d'amélioration de la reproductibilité, il sera difficile de pouvoir les utiliser comme référence. L'analyse de corrélations nous permettra ainsi de proposer des perspectives pour l'utilisation des modèles 3D comme référence de modélisation en plus des paramètres cliniques qui en sont issus.

Les analyses statistiques ont été réalisées sur StatGraph®. Pour chaque étape du processus de modélisation, la moyenne et l'écart type de la position et de l'orientation de chaque point par patient ont été calculées afin d'évaluer leur reproductibilité selon la formule de Glüer *et al.* (1995) et les recommandations de la norme ISO 5725 (équation 1.17). La précision de position (±2RMSSD) a été calculée pour l'échantillon complet (64 patients) pour le rachis, pour les thoraciques hautes seules (T1 à T6), les thoraciques basses seules (T7 à T12), les lombaires seules (L1 à L5) et des vertèbres isolées (T1, T4, T12, L1 et L5).

La variation de reproductibilité a été calculée avec le pourcentage de variation du RMSSD (équation 3.1). Si ce pourcentage est positif, il s'agit d'un gain de reproductibilité.

Une analyse de corrélation a été effectuée avec des corrélations de Pearson afin d'analyser la variation conjointe des écarts types de mesures cliniques et des écarts types de positions et d'orientations de points. Ces corrélations sont associées à des p-values obtenues avec des tests t de Student. Ces analyses nous permettront d'évaluer l'évolution des positions et des orientations et de leurs reproductibilités selon les étapes du processus de reconstruction, et de faire un parallèle avec l'évolution des paramètres cliniques.

5.3 Résultats

5.3.1 Reproductibilité de position et d'orientation des points du modèle 3D

Le tableau 5.1 montre la reproductibilité des points d'intérêts combinés ensemble pour le rachis entier, pour les thoraciques hautes (T1 à T6) pour les thoraciques basses (T7 à T12) et pour les lombaires. Il y a une amélioration de la reproductibilité des mesures entre l'étape initiale d'analyse et l'étape finale de validation, lorsque les modélisations ont suivi le processus de création de référence. Pour le rachis, la reproductibilité 3D passe de 1.3 à 1.2 mm, les reproductibilités des orientations latérales, sagittales et axiales s'améliorent respectivement de 0.3, 0.4 et 0.6 degrés. Quelles que soient les régions du rachis, nous observons une amélioration de la reproductibilité, avec des variations de gain en reproductibilité allant de 13% à 16%.

Le tableau 5.2 synthétise les reproductibilités observées pour les modélisations initiales et pour les revues d'un expert senior, SR2. SR2 retouche davantage les thoraciques hautes que les autres vertèbres (jusqu'à +11% d'évolution de reproductibilité pour les thoraciques hautes).

Si l'on regarde par vertèbres (tableau 5.3), on peut observer de légères disparités de reproductibilité, notamment pour L5 sur les orientations axiales (7,5° de reproductibilité à l'étape d'analyse, contre 4,1° à 5,5° pour les autres vertèbres). Il y a aussi une disparité sur

l'amplitude des gains en reproductibilité après l'étape de validation, avec un maximum de +43% pour la position en X de T12 et un minimum de -1% pour la position en Z de T12.

Tableau 5.1 Reproductibilité du modèle 3D par zone anatomique, selon l'étape de modélisation (reconstructions 3D initiales ou validation)

Précision de position (2RMSSD)		Position (mm)				Orientation (°)		
Zone considérée	**Expertise**	**X**	**Y**	**Z**	**3D**	**L**	**S**	**A**
Rachis	**Reconstructions 3D**	**1,8**	**1,4**	**1,3**	**1,3**	**2,4**	**2,8**	**4,3**
	Validation	**1,6**	**1,2**	**1,1**	**1,2**	**2,1**	**2,4**	**3,7**
Thoraciques Hautes	Reconstructions 3D	2,1	1,3	1,2	1,2	2,5	3,2	4,6
	Validation	1,8	1,1	1,1	1,1	2,1	2,8	3,9
Thoraciques Basses	Reconstructions 3D	1,9	1,5	1,4	1,4	2,6	2,8	4,3
	Validation	1,7	1,3	1,2	1,2	2,2	2,4	3,8
Lombaires	Reconstructions 3D	1,8	1,6	1,4	1,4	2,4	2,6	4,4
	Validation	1,5	1,4	1,2	1,2	2,1	2,2	3,7
Rachis	Gain %	14%	14%	13%	13%	14%	15%	14%
Thoraciques Hautes	Gain %	15%	14%	13%	13%	14%	15%	15%
Thoraciques Basses	Gain %	13%	14%	13%	13%	13%	15%	13%
Lombaires	Gain %	15%	15%	13%	13%	16%	16%	15%

Tableau 5.2 Reproductibilité du modèle 3D par zone anatomique, selon l'étape de modélisation (Reconstructions 3D initiales ou revues par SR2)

Précision de position (2RMSSD)		Position (mm)				Orientation (°)		
Zone considérée	**Expertise**	**A-P**	**M-L**	**P-D**	**3D**	**L**	**S**	**A**
Rachis	**Reconstructions 3D**	**1,8**	**1,4**	**1,3**	**1,3**	**2,4**	**2,8**	**4,3**
	Revue SR2	**1,8**	**1,4**	**1,3**	**1,3**	**2,4**	**2,7**	**4,1**
Thoraciques Hautes	Reconstructions 3D	2,1	1,3	1,2	1,2	2,5	3,2	4,6
	Revue SR2	2,1	1,3	1,2	1,2	2,6	2,9	4,3
Thoraciques Basses	Reconstructions 3D	1,9	1,5	1,4	1,4	2,6	2,8	4,3
	Revue SR2	1,9	1,5	1,4	1,4	2,4	2,9	4,4
Lombaires	Reconstructions 3D	1,8	1,6	1,4	1,4	2,4	2,6	4,4
	Revue SR2	1,7	1,5	1,3	1,4	2,3	2,5	4,1
Rachis	Gain %	1%	2%	1%	1%	1%	4%	3%
Thoraciques Hautes	Gain %	1%	1%	0%	1%	-4%	11%	5%
Thoraciques Basses	Gain %	0%	1%	1%	3%	5%	-4%	-1%
Lombaires	Gain %	4%	5%	2%	2%	6%	5%	7%

La figure 5.3 montre l'évolution conjointe de la reproductibilité pour les paramètres cliniques et pour les positions et orientations du rachis dans sa globalité entre l'étape de reconstructions 3D initiales, l'étape de revue par SR2 et l'étape de validation. On peut observer que les améliorations de reproductibilité sont variables selon les paramètres et selon l'étape. SR2 par exemple semble beaucoup améliorer la reproductibilité de L1/S1 et de la rotation axiale de L5, mais semble aussi avoir moins d'impact dans ses corrections sur les rotations axiales de T1 et de T4.

Tableau 5.3 Évolution de la reproductibilité de position et d'orientation de vertèbres isolées entre l'étape de reconstructions 3D initiales et l'étape de validation

Précision de position (2RMSSD)		Position (mm)				Orientation (°)		
Zone	Expertise	**X**	**Y**	**Z**	**3D**	**L**	**S**	**A**
T1	Reconstructions 3D	1,3	1,5	1,1	1,1	2,7	3,5	5,5
	Validation	1,1	1,2	1,0	1,0	2,2	3,0	4,7
T4	Reconstructions 3D	2,5	1,4	1,3	1,3	2,6	3,4	4,7
	Validation	2,1	1,2	1,1	1,1	2,3	2,9	3,9
T12	Reconstructions 3D	1,9	1,6	1,6	1,7	2,8	3,0	5,3
	Validation	1,6	1,4	1,4	1,4	2,4	2,5	4,5
L1	Reconstructions 3D	1,6	1,4	1,3	1,4	2,6	2,8	4,1
	Validation	1,4	1,2	1,2	1,2	2,2	2,3	3,6
L5	Reconstructions 3D	2,5	2,7	1,7	1,8	4,2	3,4	7,5
	Validation	2,0	2,2	1,5	1,6	3,6	2,8	6,0
T1	Gain %	15%	15%	13%	13%	18%	15%	15%
T4	Gain %	16%	13%	13%	13%	14%	14%	17%
T12	Gain %	13%	14%	14%	14%	14%	17%	14%
L1	Gain %	14%	14%	13%	13%	14%	20%	12%
L5	Gain %	18%	18%	13%	13%	14%	15%	20%

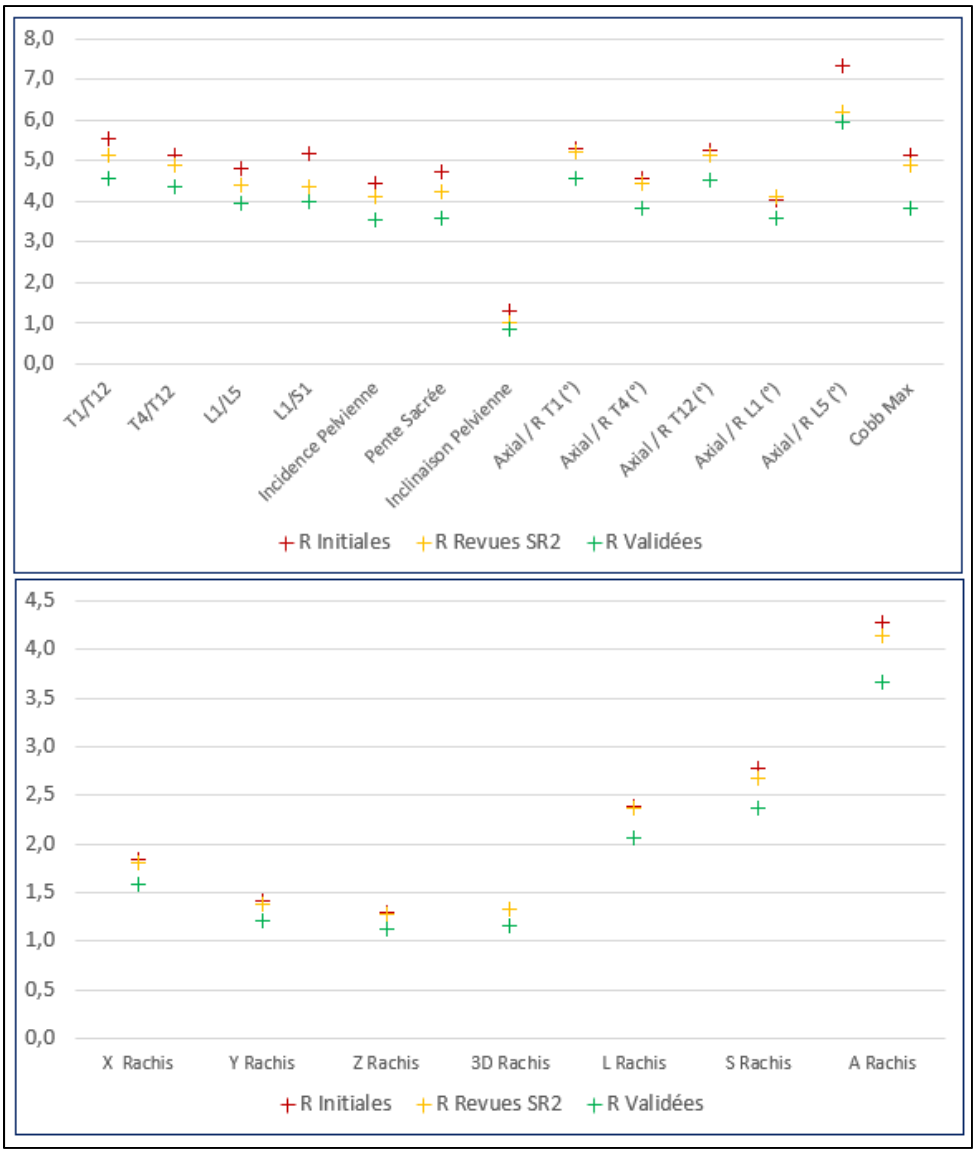


Figure 5.3 Évolution conjointe des reproductibilités (2RMSSD) des paramètres cliniques (en haut) et des positions et orientations du rachis (en bas)

La figure 5.4 montre l'évolution de reproductibilité pour la position et l'orientation globale du rachis, entre l'étape d'analyse, l'étape de revue par SR2 et l'étape de validation, pour deux catégories de patients, ceux présentant une bonne visibilité sur l'image frontale (visibilité AP), et ceux présentant une mauvaise visibilité sur l'image frontale). On peut observer une amélioration progressive de la reproductibilité au fur et à mesure du processus de production de modélisation, avec pour les patients à mauvaise visibilité, des reproductibilités initiales

moins bonnes que pour les autres (comme pour les mesures cliniques cependant, aucune différence statistiquement significative n'a été observée).

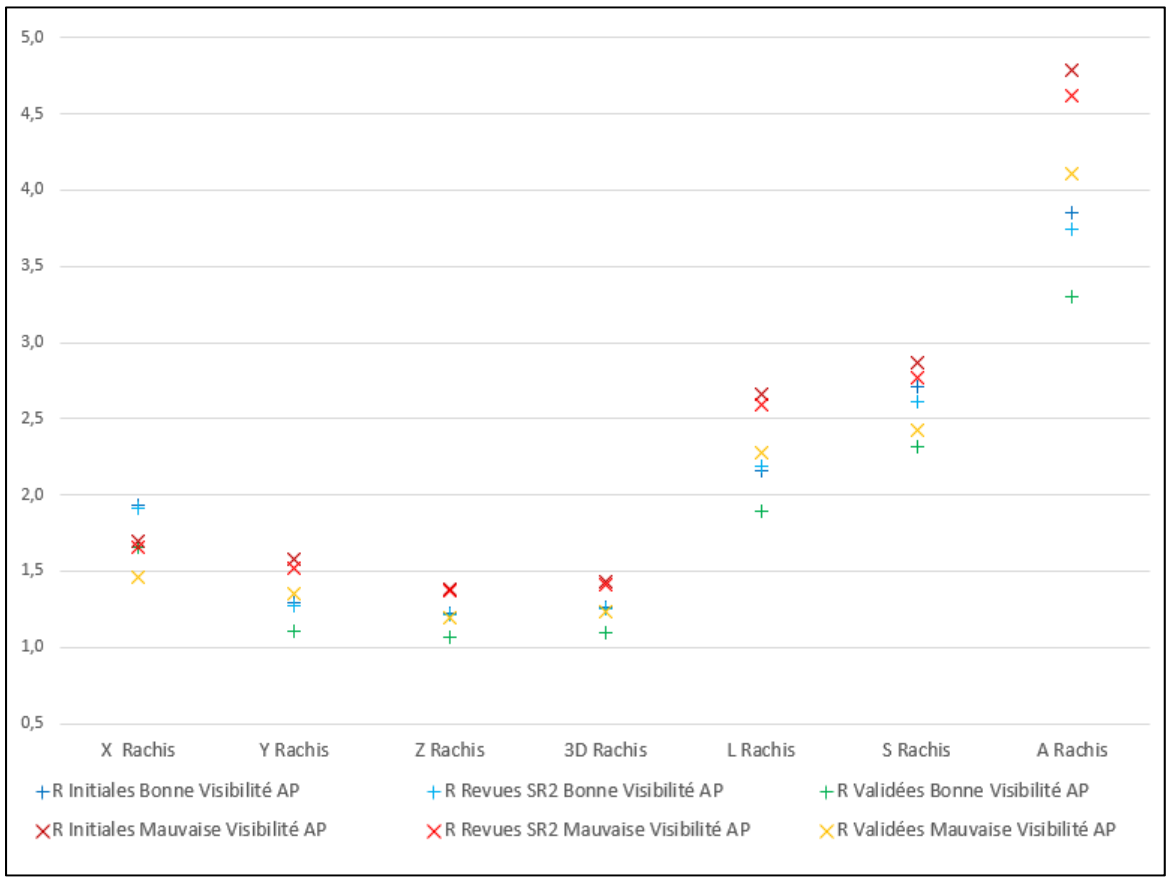


Figure 5.4 Évolution de la reproductibilité de position et d'orientation du rachis en fonction des étapes de modélisations, pour les patients à mauvaise visibilité en vue frontale (AP) et les patients à bonne visibilité en vue frontale (AP)

5.3.2 Corrélations avec la dispersion des paramètres cliniques

Le tableau 5.4 montre les corrélations de Pearson entre les écarts types observés pour la position et l'orientation des ensembles de points de T1, T4, T12, L1 et L5, et les écarts types observés pour les mesures de balance sagittale pour les modélisations indépendantes de l'étape d'analyse. On peut voir qu'il existe des corrélations significatives entre la dispersion de certaines mesures et la dispersion de certains points anatomiques. Pour la cyphose T1/T12, il existe une corrélation positive et significative entre les écarts types des positions des points de T1 et les écarts types de mesures de T1/12 (valeur de p situées entre 0,0058 et 0,0247). Pour T4/T12, les écarts types des orientations sagittales de T4 et T12 sont corrélés

aux écarts types de T4/12 (valeurs de p respectives de 0,0007 et 0,0002). Pour L1/L5, les écarts types de l'orientation sagittale de L1 sont corrélés aux écarts types de L1/L5 (valeur de p de 0,0000) mais pas l'orientation sagittale de L5 (valeur de p de 0,5683). Globalement, les corrélations observées sont modérées, les coefficients de corrélations étant situés entre 0,28 et 0,58 pour les corrélations significatives.

Tableau 5.4 Coefficients de corrélations de Pearson (CCP) et valeurs de p (p) après test *t* de Student, entre la dispersion des paramètres cliniques de mesure de balance sagittale et la dispersion de la position ou de l'orientation des vertèbres associées à ces paramètres cliniques à l'étape d'analyse seule

		T1 3D	**T1 A**	**T1 L**	**T1 S**	**T1 X**	**T1 Y**	**T1 Z**
T1/T12	**CCP**	**0,29**	0,23	0,16	**0,33**	**0,34**	**0,32**	**0,28**
	p	**(0,0209)**	(0,0678)	(0,1961)	**(0,0071)**	**(0,0058)**	**(0,01)**	**(0,0247)**
		T4 3D	**T4 A**	**T4 L**	**T4 S**	**T4 X**	**T4 Y**	**T4 Z**
T4/T12	**CCP**	0,24	**0,28**	0,15	**0,41**	0,15	0,15	0,23
	p	(0,0528)	**(0,0228)**	(0,2221)	**(0,0007)**	(0,2443)	(0,246)	(0,0678)
		L1 3D	**L1 A**	**L1 L**	**L1 S**	**L1 X**	**L1 Y**	**L1 Z**
L1/L5	**CCP**	0,17	**0,33**	-0,15	**0,58**	0,09	0,18	0,14
	p	(0,1731)	**(0,0085**)	(0,2224)	**(0,0000)**	(0,462)	(0,165)	(0,2851)
		T12 3D	**T12 A**	**T12 L**	**T12 S**	**T12 X**	**T12 Y**	**T12 Z**
T1/T12	**CCP**	0,17	0,07	0,05	**0,35**	**0,30**	0,17	0,18
	p	(0,178)	(0,6000)	(0,7098)	**(0,0041)**	**(0,0166)**	(0,18)	(0,1438)
		T12 3D	**T12 A**	**T12 L**	**T12 S**	**T12 X**	**T12 Y**	**T12 Z**
T4/T12	**CCP**	0,09	0,17	**0,28**	**0,45**	0,21	0,12	0,12
	p	(0,4929)	(0,1767)	**(0,0269)**	**(0,0002)**	(0,0983)	(0,3582)	(0,3314)
		L5 3D	**L5 A**	**L5 L**	**L5 S**	**L5 X**	**L5 Y**	**L5 Z**
L1/L5	**CCP**	-0,10	-0,03	0,06	-0,07	-0,06	-0,00	-0,11
	p	(0,426)	(0,8345)	(0,6544)	(0,5683)	(0,6272)	(0,9988)	(0,3786)

Le tableau 5.5 montre les corrélations entre les écarts types des orientations axiales de T1, T4, T12, L1 et L5 et les écarts types des rotations axiales des mêmes vertèbres qui sont les paramètres cliniques issus du logiciel commercial. On peut observer les corrélations attendues entre les dispersions d'orientation et de rotation pour chaque vertèbre individuelle (valeurs de p de 0,0000). Les corrélations significatives attendues sont très fortes, avec des coefficients de corrélation de Pearson de 0,93 à 0,99.

Tableau 5.5 Coefficients de corrélations de Pearson (CCP) et valeurs de p (p) après test *t* de Student, entre la dispersion des paramètres cliniques de rotation axiale de T1, T4, T12, L1 et L5 et la dispersion des mesures d'orientations dans le plan axial de ces vertèbres à l'étape d'analyse seule

		Axial / R T1	Axial / R T4	Axial / R T12	Axial / R L1	Axial / R L5
A T1	**CCP** **p**	**0,93** **(0,0000)**	0,34 (0,5923)	0,11 (0,3719)	0,25 (0,0451)	0,15 (0,2481)
A T4	**CCP** **p**	0,03 (0,0413)	**0,97** **(0,0000)**	-0,03 (0,8066)	0,02 (0,8899)	0,0861 (0,4990)
A T12	**CCP** **p**	-0,01 (0,9588)	-0,04 (0,7516)	**0,96** **(0,0000)**	**0,55** **(0,0000)**	0,02 (0,8881)
A L1	**CCP** **p**	0,03 (0,8174)	0,02 (0,8679)	**0,58** **(0,0000)**	**0,97** **(0,0000)**	0,08 (0,5460)
A L5	**CCP** **p**	0,09 (0,4973)	0,07 (0,5923)	-0,04 (0,7307)	0,08 (0,5292)	**0,99** **(0,0000)**

5.3.3 Corrélation avec les variations de reproductibilité

Le tableau 5.6 montre les mêmes corrélations mais cette fois en regard des différences d'écarts types entre l'étape d'analyse et l'étape de validation pour la mesure T1/T12 et la précision de position et d'orientation de T1 et T12. La variation de dispersion de la mesure T1/T12 est corrélée à la variation de dispersion des positions 3D, en Y et en Z et de l'orientation sagittale de T1 (valeurs de p comprises entre 0,0004 et 0,0189). Cela veut dire qu'il y a un lien entre les deux variations de reproductibilité, et pas seulement un lien entre les deux reproductibilités initiales comme présenté au tableau 5.4. Les corrélations significatives pour T1 sont cependant modérées avec des coefficients de Pearson de 0,29 à 0,43. Nous voyons que pour T12 il n'y a pas de corrélations observées entre la mesure T1/T12 et les positions et orientations de la vertèbre T12.

5.3.4 Discussion

On peut voir que comme pour les paramètres cliniques, il y a une amélioration de la reproductibilité lorsque la méthode de création de référence est appliquée, quelles que soient les régions du rachis (tableau 5.1).

Tableau 5.6 Coefficients de corrélations de Pearson (CCP) et valeurs de p (p) après test *t* de Student, entre l'augmentation de reproductibilité de T1/T12 et l'augmentation de reproductibilité de position et d'orientation de T1 et T12

		T1 3D	T1 A	T1 L	T1 S	T1 X	T1 Y	T1 Z
T1/T12	**CCP**	**0,31**	0,1265	0,22	**0,37**	0,21	**0,43**	**0,29**
	p	**(0,0122)**	(0,3193)	(0,0782)	**(0,0030)**	(0,1004)	**(0,0004)**	**(0,0189)**
		T12 3D	T12 A	T12 L	T12 S	T12 X	T12 Y	T12 Z
T1/T12	**CCP**	-0,03	0,05	0,13	0,2341	0,20	0,04	-0,00
	p	(0,8301)	(0,7277)	(0,3153)	(0,0626)	(0,114)	(0,7256)	(0,9818)

Nous avons vu également avec les modélisations revues par SR2 qu'un expert va regarder certaines positions de points d'intérêts plus que d'autres, comme ce fut le cas pour certaines mesures cliniques (tableau 5.2 et figure 5.3). Cela rejoint nos constats du chapitre 3 sur la différence de comportement des correcteurs, avec des divergences observables et quantifiables sur ce qu'ils corrigent et les structures sur lesquelles ils portent leur attention préférentiellement.

Nous avons vu également que les augmentations de reproductibilités divergent selon les catégories de patients, de même que le comportement des experts et la confiance qu'ils donnent à leurs mesures (figure 5.5). Une analyse approfondie de chacune des relations vertèbre-mesure clinique permettrait de cartographier quel point (position ou orientation) est le plus sollicité pour chaque ajustement de mesures cliniques lors du processus de reconstruction 3D.

Les analyses effectuées permettent d'apporter des preuves sur le lien entre la reproductibilité des paramètres cliniques et la reproductibilité de la position et de l'orientation des points du modèle 3D (tableau 5.4, 5.5 et 5.6). S'il est rassurant de constater qu'il y a bien des corrélations significatives entre la dispersion de certains points et la dispersion des paramètres cliniques auxquels ils sont liés, une analyse plus approfondie serait nécessaire devant la multiplicité des paramètres cliniques et la multiplicité des points.

5.4 Résumé

La reconstruction 3D de colonne vertébrale à partir d'images bi-planaires est une tâche d'analyse d'image pour laquelle il n'existe pas d'étalon or, à cause de la position debout. Pour rappel, ce sont les paramètres cliniques calculés automatiquement sur les modèles 3D qui ont reçu une validation pour un usage clinique.

Le travail exploratoire du chapitre 5 nous permet de répondre à des questions assez simples, pour ouvrir la voie à des analyses plus approfondies. Comme vu au chapitre 3, il est possible à l'aide d'experts et d'une méthodologie de reconstruction incluant des discussions d'obtenir des références de mesures cliniques plus fiables pour obtenir une réalité terrain. De ces analyses sur les mesures il était donc intéressant d'observer l'évolution de la fiabilité des modèles 3D en parallèle de celle des mesures cliniques. Nous avons ainsi pu voir que non seulement les dispersions des mesures cliniques dans l'échantillon sont très similaires à celles des points des modèles 3D à l'étape d'analyse du processus de reconstruction 3D. De plus, l'augmentation de la reproductibilité entre l'étape d'analyse et l'étape de validation est également corrélée à l'augmentation de reproductibilité de la position et de l'orientation du modèle 3D, pour certains couples vertèbre-mesure clinique.

Ces résultats préliminaires posent les bases d'investigation d'un travail beaucoup plus large. En effet, pour pouvoir inclure les modèles 3D en tant que référence, il faudrait être capable, dans une méthode d'évaluation de performance telle que celle mise en place au chapitre 4, de formaliser les liens entre chacun des paramètres cliniques et chacun des points d'intérêts du modèle. Une fonctionnalité intéressante dans cette idée serait, en cas d'erreur de mesures cliniques, de pouvoir faire un lien direct avec les points concernés pour identifier quel point a été mal identifié. Avec la méthode mise en place, il y aurait en parallèle la caractérisation de la base de données d'évaluation qui donnerait en outre l'information sur le type d'image où l'erreur de position ou d'orientation a été effectuée, afin de mieux guider les interprétations des résultats d'évaluations.

Il existe cependant un biais par rapport à la relation entre les experts et les modèles 3D, c'est que la formation sur le logiciel commercial qu'ils utilisent tous les jours est ciblée sur la réussite des mesures cliniques davantage que sur la réussite du modèle. Il y aurait donc une réflexion à apporter sur ce biais si l'on veut pouvoir utiliser les modèles 3D comme référence.

CONCLUSION ET RECOMMANDATIONS

L'objectif initial de ce travail était d'améliorer les méthodes d'évaluations en donnant aux développeurs une méthode d'évaluation pour visualiser directement la performance de leurs méthodes de reconstructions 3D, et de construire un lien entre les développeurs et les cliniciens en ajoutant une sémantique clinique à cette méthode d'évaluation. Dans le contexte des reconstructions 3D de colonne vertébrale à partir d'images bi-planaires, il s'agissait donc de résoudre plusieurs problématiques parmi lesquelles l'absence d'étalon or et l'absence de critères cliniques d'évaluation.

La littérature regorge d'études qui, n'ayant pas à disposition d'étalon or, se sont appuyées sur des experts pour construire leur référence. Maier-Hein *et al.* (2018) a cependant mis en évidence que les pratiques à ce niveau-là étaient hétérogènes et manquaient fortement de standardisation ou de pratiques communes. En s'appuyant sur les expériences de terrain de l'entreprise partenaire, nous avons investigué l'intérêt de la discussion entre experts pour la mise en place d'une méthodologie de production de référence plus reproductible. Il est apparu qu'effectivement, inclure des contre-vérifications et des validations entre experts permet d'améliorer la reproductibilité des mesures cliniques dans notre domaine d'application. La réalité terrain que nous proposons d'obtenir avec cette nouvelle méthode de création de référence (figure 3.2) sont des intervalles de confiances de mesures cliniques autour d'une valeur moyenne, issues des reconstructions 3D de colonne vertébrale. Cette méthode nécessite la participation de plusieurs experts, pour la première analyse, pour la revue et pour la validation des reconstructions 3D. Dans notre proposition, 5 experts au total ont été sollicités pour construire la référence, or dans la littérature, jamais un nombre optimal d'expert n'a été évoqué pour construire des références. Pour ce qui est de la multiplicité des profils d'experts, un biais existe dans notre étude dans la mesure ou chacun des experts ont suivi la même formation initiale et ont plusieurs années d'expériences en reconstruction 3D de colonne vertébrale. Cependant, malgré le manque de diversité en termes de formations, nous avons mis en évidence que chacun a apporté de l'expertise différente pour augmenter la confiance dans les mesures. Il serait donc intéressant de davantage explorer l'impact de la diversité des profils dans la variation de la confiance dans les mesures. Le consensus pour

créer une référence n'a pas été investigué à notre connaissance dans la littérature, et nous avons apporté des preuves que l'inclusion de discussions entre experts sur le processus proposé permet d'obtenir des intervalles de confiances de mesures plus étroits, et donc des références plus fiables qui peuvent être considérées comme des réalités terrains (Cardoso *et al.*, 2014).

Nous avons mis en place un score de visibilité image afin de mieux caractériser la base de données d'évaluations, afin de donner le maximum d'informations sur sa représentativité, que ce soit en termes d'âge, de sévérité des déformations, ou de la visibilité des structures dans les images. Pour ce dernier point, nous avons par ailleurs proposé un score de visibilité image qui permette d'inclure à la fois la qualité physique des images et les déformations pathologiques, qui sont deux caractéristiques ayant de potentiels impacts sur la visibilité des structures.

Parmi les mesures de performance, nous avons bien évidemment inclus des métriques pour quantifier les erreurs de mesures. Puisque notre méthode d'évaluation est centrée sur les mesures cliniques issues des reconstructions 3D de colonne vertébrale à partir d'images EOS, nous avons sélectionné des métriques liées aux mesures dans l'image, soit le taux de succès, le pourcentage d'erreur SMAPE, la différence moyenne MAD, et le pourcentage d'accord expert. Ce dernier point est fortement lié aux critères de performances cliniques que nous avons également mis en place. Une revue de littérature clinique nous a en effet permis d'adapter des critères de gravités selon les erreurs rencontrées sur certaines mesures cliniques, afin de hiérarchiser ces erreurs selon leur gravité, et donc leur potentiel impact sur une prise en charge thérapeutique ou diagnostique. Ces critères cliniques de performance n'ont jusqu'ici jamais été mis en place dans la littérature, à notre connaissance, pour l'évaluation de performance de méthodes d'analyse d'image. Maier-Hein *et al.* (2018) a pourtant recommandé de se rapprocher des sociétés savantes pour améliorer les méthodes d'évaluation, et dans cette optique nous nous sommes fortement inspirés des recommandations officielles de la SRS et de la SOSORT pour construire nos critères cliniques. Il est cependant important de repréciser qu'ajouter des critères cliniques n'implique pas de donner un pouvoir diagnostique aux méthodes que l'on évalue, c'est

simplement un guide pour les développeurs pour leur permettre de hiérarchiser les erreurs selon la gravité clinique qu'elles impliquent.

La construction de la méthode d'évaluation s'est donc faite en deux étapes, avec d'abord la mise en place de la référence et ensuite l'ajout des critères cliniques d'évaluation. Nous avons également inclus des métriques de quantification des erreurs sur appui de la littérature, qui permettent de facilement repositionner la performance de la méthode évaluée. Nous avons ainsi pu conduire plusieurs évaluations dont celle d'un algorithme automatique de reconstruction 3D de colonne vertébrale. Cet algorithme est beaucoup plus performant pour la pédiatrie que pour les adultes, ce qui va permettre aux développeurs de mieux comprendre comment améliorer leur méthode de reconstruction. De plus, la méthode pour caractériser la visibilité des images couplée à la méthode de production de référence rend facile l'intégration de nouveaux sujets dans la base de données d'évaluation, afin de renforcer sa représentativité. Outre les algorithmes, cette méthode est également utilisable pour l'évaluation de performance des experts, pendant une formation initiale ou dans le cadre d'une formation continue. Il serait dans cette idée intéressant de réfléchir à des objectifs de performance à atteindre, au regard de performances atteintes par les experts seniors par exemple. Les évaluations permettent aussi, grâce aux critères cliniques et aux tables d'impacts de donner de bons indicateurs de développement ou de formation.

La retombée majeure de ce projet est d'avoir proposé une méthode d'évaluation qui prenne en compte des exigences cliniques, et qui soit généralisable à d'autres tâches. La méthodologie est en effet transposable facilement, car la littérature clinique ne manque pas de recommandations de bonnes pratiques quant à la classification de pathologies, quelle que soit le domaine d'imagerie. La mobilisation des experts est quasi systématique dans la littérature pour construire des références en l'absence d'étalon or, nous proposons une méthode de construction en appliquant des étapes de corrections et de validations entre les experts pour isoler les images très difficiles et arriver à un consensus, et ce tout en augmentant la reproductibilité des mesures. Nous recommandons dans ce sens le concours de plusieurs experts pour la création de référence, afin de renforcer la reproductibilité des mesures.

Nous suggérons pour la poursuite des travaux de rester dans le domaine d'application des reconstructions 3D de la colonne vertébrale, et d'investiguer les liens entre le modèle et les paramètres cliniques afin de compléter l'évaluation de performance avec des informations quantitatives sur les erreurs dans les modèles. Notre travail au chapitre 5 a permis en effet de mettre en évidence certains liens, qu'il serait très utile d'analyser plus finement pour améliorer la méthode d'évaluation.

Finalement, ce travail a consisté à améliorer les évaluations des méthodes de reconstructions 3D de colonne vertébrale. Dans ce contexte, il faut rappeler l'importance de ces évaluations dans un domaine qui voit l'automatisation prendre de plus en plus de place. L'automatisation consiste à être moins dépendant des experts, qui réalisent des tâches longues et fastidieuses pour analyser les images, avec des erreurs humaines qui peuvent apparaître. Améliorer les méthodes d'évaluation en sollicitant les experts en amont, de façon plus pertinente et plus efficace pour construire les références d'évaluation, constitue la retombée majeure de notre travail.

ANNEXE I

INFORMATIONS COMPLÉMENTAIRES SUR LES DONNÉES UTILISÉES AUX CHAPITRES 3, 4 ET 5

Le tableau-A I-1 synthétise quelques-unes des caractéristiques des sous échantillons, dont la répartition des scores de visibilité et la répartition des angles de Cobb maximal.

Tableau-A I-1 Caractéristiques et effectifs des données utilisées dans ce travail

Catégorie	Définition	Caractéristiques	n
Tous	56F 8M	Âge moyen : 34.2y ± 25.6y	64
Pédiatrie	Patient n'ayant pas 18 ans	Âge moyen : 14y ± 1y	35
Adultes	Patient ayant passé 18 ans	Âge moyen : 56y ±21y	29
Léger/modéré	Patient n'ayant aucun angle de Cobb supérieur à 40 degrés	Cobb max : 23.11° ± 11.1°	20
Sévère	Patient ayant au moins un angle de Cobb supérieur à 40 degrés	Cobb max : 57.6° ± 12.3°	44
Bonne Visibilité AP	Patients ayant reçu un score de visibilité des structures en AP de plus de 50%	Score moyen : 57.4% ± 4.8% Cobb max 45.3° ±22.6°	37
Mauvaise Visibilité AP	Patients ayant reçu un score de visibilité des structures en AP de moins de 50%	Score moyen : 40.62% ± 7.3% Cobb max 49.8° ± 15.7°	27
Bonne Visibilité LAT	Patients ayant reçu un score de visibilité des structures en LAT de plus de 50%	Score moyen : 63.6% ± 9.2% Cobb max 44.5 ± 19.7°	52
Mauvaise Visibilité LAT	Patients ayant reçu un score de visibilité des structures en LAT de moins de 50%	Score Moyen : 42.0% ± 7.9% Cobb max 58.8° ± 16.8°	12
Bonne Visibilité Thoraciques	Patients ayant reçu un score de visibilité des structures thoraciques de plus de 50%	Score Moyen : 60.6% ± 8.3%	11
Mauvaise Visibilité Thoraciques	Patients ayant reçu un score de visibilité des structures thoraciques de moins de 50%	Score Moyen : 36.3% ± 7%	53
Bonne Visibilité Lombaires	Patients ayant reçu un score de visibilité des structures lombaires de plus de 50%	Score Moyen : 56.2% ± 5.2%	35
Mauvaise Visibilité Lombaires	Patients ayant reçu un score de visibilité des structures lombaires de moins de 50%	Score Moyen : 39.3% ± 7.1%	29

BIBLIOGRAPHIE

Akhondi-Asl, A., & Warfield, S. (2019). *A Tutorial Introduction to STAPLE*.

Aubert, B., Vazquez, C., Cresson, T., Parent, S., & de Guise, J. A. (2019). Toward Automated 3D Spine Reconstruction from Biplanar Radiographs Using CNN for Statistical Spine Model Fitting. *IEEE Transactions on Medical Imaging*, *38*(12), 2796-2806. https://doi.org/10.1109/TMI.2019.2914400

Aubert, B., Vergari, C., Ilharreborde, B., Courvoisier, A., & Skalli, W. (2016). 3D reconstruction of rib cage geometry from biplanar radiographs using a statistical parametric model approach. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *4*(5), 281-295. https://doi.org/10.1080/21681163.2014.913990

Bettany-Saltikov, J., Turnbull, D., Ng, S. Y., & Webb, R. (2017). Management of Spinal Deformities and Evidence of Treatment Effectiveness. *The Open Orthopaedics Journal*, *11*, 1521-1547. https://doi.org/10.2174/1874325001711011521

Cárdenes, R., de Luis-García, R., & Bach-Cuadra, M. (2009). A multidimensional segmentation evaluation for medical image data. *Computer Methods and Programs in Biomedicine*, *96*(2), 108-124. https://doi.org/10.1016/j.cmpb.2009.04.009

Cardoso, J. R., Pereira, L. M., Iversen, M. D., & Ramos, A. L. (2014). What is gold standard and what is ground truth? *Dental Press Journal of Orthodontics*, *19*(5), 27-30. https://doi.org/10.1590/2176-9451.19.5.027-030.ebo

Chung, N., Cheng, Y.-H., Po, H.-L., Ng, W.-K., Cheung, K.-C., Yung, H.-Y., & Lai, Y.-M. (2018). Spinal phantom comparability study of Cobb angle measurement of scoliosis using digital radiographic imaging. *Journal of Orthopaedic Translation*, *15*, 81-90. https://doi.org/10.1016/j.jot.2018.09.005

Deschênes, S., Charron, G., Beaudoin, G., Labelle, H., Dubois, J., Miron, M.-C., & Parent, S. (2010). Diagnostic Imaging of Spinal Deformities : Reducing Patients Radiation Dose With a New Slot-Scanning X-ray Imager. *Spine*, *35*(9), 989-994. https://doi.org/10.1097/BRS.0b013e3181bdcaa4

Despotović, I., Goossens, B., & Philips, W. (2015). MRI Segmentation of the Human Brain : Challenges, Methods, and Applications. *Computational and Mathematical Methods in Medicine*, *2015*. https://doi.org/10.1155/2015/450341

Dewalle-Vignion, A.-S., Betrouni, N., Baillet, C., & Vermandel, M. (2015). Is STAPLE algorithm confident to assess segmentation methods in PET imaging? *Physics in Medicine and Biology*, *60*(24), 9473-9491. https://doi.org/10.1088/0031-9155/60/24/9473

Galbusera, F., Niemeyer, F., Wilke, H.-J., Bassani, T., Casaroli, G., Anania, C., Costa, F., Brayda-Bruno, M., & Sconfienza, L. M. (2019). Fully automated radiological analysis of spinal disorders and deformities : A deep learning approach. *European Spine Journal: Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, *28*(5), 951-960. https://doi.org/10.1007/s00586-019-05944-z

Glüer, C.-C., Blake, G., Lu, Y., Blunt1, B. A., Jergas1, M., & Genant1, H. K. (1995). Accurate assessment of precision errors : How to measure the reproducibility of bone densitometry techniques. *Osteoporosis International*, *5*(4), 262-270. https://doi.org/10.1007/BF01774016

Heimann, T., Ginneken, B. van, Styner, M. A., Arzhaeva, Y., Aurich, V., Bauer, C., Beck, A., Becker, C., Beichel, R., Bekes, G., Bello, F., Binnig, G., Bischof, H., Bornik, A., Cashman, P. M. M., Chi, Y., Cordova, A., Dawant, B. M., Fidrich, M., … Wolf, I. (2009). Comparison and Evaluation of Methods for Liver Segmentation From CT Datasets. *IEEE Transactions on Medical Imaging*, *28*(8), 1251-1265. https://doi.org/10.1109/TMI.2009.2013851

Humbert, L., De Guise, J. A., Aubert, B., Godbout, B., & Skalli, W. (2009). 3D reconstruction of the spine from biplanar X-rays using parametric models based on transversal and longitudinal inferences. *Medical Engineering & Physics*, *31*(6), 681-687. https://doi.org/10.1016/j.medengphy.2009.01.003

Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*(9), 850-863. https://doi.org/10.1109/34.232073

Ilharreborde, B., Ferrero, E., Alison, M., & Mazda, K. (2016). EOS microdose protocol for the radiological follow-up of adolescent idiopathic scoliosis. *European Spine Journal*, *25*(2), 526-531. https://doi.org/10.1007/s00586-015-3960-8

Ilharreborde, B., Steffen, J. S., Nectoux, E., Vital, J. M., Mazda, K., Skalli, W., & Obeid, I. (2011). Angle measurement reproducibility using EOS three-dimensional reconstructions in adolescent idiopathic scoliosis treated by posterior instrumentation. *Spine*, *36*(20), E1306-1313. https://doi.org/10.1097/BRS.0b013e3182293548

ISO 5725-1:1994. (s. d.). Consulté 15 juin 2021, à l'adresse https://www.iso.org/obp/ui/#iso:std:iso:5725:-1:ed-1:v1:en

Jannin, P., Fitzpatrick, J. M., Hawkes, D. J., Pennec, X., Shahidl, R., & Vannier, M. W. (2002). Validation of medical image processing in image-guided therapy. *IEEE Transactions on Medical Imaging*, *21*(12), 1445-1449. https://doi.org/10.1109/TMI.2002.806568

Jirot, A., Hocquelet, A., Reaux, C., & De Seze, M. (2015). A comparative study between two different 3D reconstruction methods by bi-planar radiographic in upright posture : Biomod 3sand sterEOS®. *Annals of Physical and Rehabilitation Medicine*, *58*, e89. https://doi.org/10.1016/j.rehab.2015.07.217

Kim, K. C., Yun, H. S., Kim, S., & Seo, J. K. (2020). Automation of Spine Curve Assessment in Frontal Radiographs Using Deep Learning of Vertebral-Tilt Vector. *IEEE Access*, *8*, 84618-84630. https://doi.org/10.1109/ACCESS.2020.2992081

Korez, R., Putzier, M., & Vrtovec, T. (2020). A deep learning tool for fully automated measurements of sagittal spinopelvic balance from X-ray images : Performance evaluation. *European Spine Journal: Official Publication of the European Spine Society, the European Spinal Deformity Society, and the European Section of the Cervical Spine Research Society*, *29*(9), 2295-2305. https://doi.org/10.1007/s00586-020-06406-7

Laurent, P., Cresson, T., Vázquez, C., Hagemeister, N., & Guise, J. A. de. (2016). A multi-criteria evaluation platform for segmentation algorithms. *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 6441-6444. https://doi.org/10.1109/EMBC.2016.7592203

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A. P., Carass, A., Feldmann, C., Frangi, A. F., Full, P. M., van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B. A., März, K., … Kopp-Schneider, A. (2018). Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, *9*(1), 5217. https://doi.org/10.1038/s41467-018-07619-7

Melhem, E., Assi, A., El Rachkidi, R., & Ghanem, I. (2016). EOS(®) biplanar X-ray imaging : Concept, developments, benefits, and limitations. *Journal of Children's Orthopaedics*, *10*(1), 1-14. https://doi.org/10.1007/s11832-016-0713-0

Negrini, S., Donzelli, S., Aulisa, A. G., Czaprowski, D., Schreiber, S., de Mauroy, J. C., Diers, H., Grivas, T. B., Knott, P., Kotwicki, T., Lebel, A., Marti, C., Maruyama, T., O'Brien, J., Price, N., Parent, E., Rigo, M., Romano, M., Stikeleather, L., … Zaina, F. (2018). 2016 SOSORT guidelines : Orthopaedic and rehabilitation treatment of idiopathic scoliosis during growth. *Scoliosis and Spinal Disorders*, *13*. https://doi.org/10.1186/s13013-017-0145-8

Nérot, A., Choisne, J., Amabile, C., Travert, C., Pillet, H., Wang, X., & Skalli, W. (2015). A 3D reconstruction method of the body envelope from biplanar X-rays : Evaluation of its accuracy and reliability. *Journal of Biomechanics*, *48*(16), 4322-4326. https://doi.org/10.1016/j.jbiomech.2015.10.044

O'Brien, M. F. & Spinal Deformity Study Group. (2004). *Radiographic Measurement Manual*. Medtronic Sofamor Danek USA. https://books.google.ca/books?id=If_0swEACAAJ

Rehm, J., Germann, T., Akbar, M., Pepke, W., Kauczor, H.-U., Weber, M.-A., & Spira, D. (2017). 3D-modeling of the spine using EOS imaging system : Inter-reader reproducibility and reliability. *PLOS ONE*, *12*(2), e0171258. https://doi.org/10.1371/journal.pone.0171258

Schwab, F., Ungar, B., Blondel, B., Buchowski, J., Coe, J., Deinlein, D., DeWald, C., Mehdian, H., Shaffrey, C., Tribus, C., & Lafage, V. (2012). Scoliosis Research Society-Schwab adult spinal deformity classification : A validation study. *Spine*, *37*(12), 1077-1082. https://doi.org/10.1097/BRS.0b013e31823e15e2

Şerbănescu, M. S., Oancea, C. N., Streba, C. T., Pleşea, I. E., Pirici, D., Streba, L., & Pleşea, R. M. (2020). Agreement of two pre-trained deep-learning neural networks built with transfer learning with six pathologists on 6000 patches of prostate cancer from Gleason2019 Challenge. *Romanian Journal of Morphology and Embryology = Revue Roumaine De Morphologie Et Embryologie*, *61*(2), 513-519. https://doi.org/10.47162/RJME.61.2.21

Skalli, W., Lavaste, F., & Descrimes, J. L. (1995). Quantification of three-dimensional vertebral rotations in scoliosis : What are the true values? *Spine*, *20*(5), 546-553. https://doi.org/10.1097/00007632-199503010-00008

Slattery, C., & Verma, K. (2018). Classifications in Brief : The Lenke Classification for Adolescent Idiopathic Scoliosis. *Clinical Orthopaedics and Related Research*, *476*(11), 2271-2276. https://doi.org/10.1097/CORR.0000000000000405

Somoskeöy, S., Tunyogi-Csapó, M., Bogyó, C., & Illés, T. (2012). Clinical validation of coronal and sagittal spinal curve measurements based on three-dimensional vertebra vector parameters. *The Spine Journal: Official Journal of the North American Spine Society*, *12*(10), 960-968. https://doi.org/10.1016/j.spinee.2012.08.175

Stokes, I. A. (1994). Three-dimensional terminology of spinal deformity. A report presented to the Scoliosis Research Society by the Scoliosis Research Society Working Group on 3-D terminology of spinal deformity. *Spine*, *19*(2), 236-248.

Stute, S., Carlier, T., Cristina, K., Noblet, C., Martineau, A., Hutton, B., Barnden, L., & Buvat, I. (2011). Monte Carlo simulations of clinical PET and SPECT scans : Impact of the input data on the simulated images. *Physics in Medicine and Biology*, *56*(19), 6441-6457. https://doi.org/10.1088/0031-9155/56/19/017

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation : Analysis, selection, and tool. *BMC Medical Imaging*, *15*, 29. https://doi.org/10.1186/s12880-015-0068-x

Terran, J., Schwab, F., Shaffrey, C. I., Smith, J. S., Devos, P., Ames, C. P., Fu, K.-M. G., Burton, D., Hostin, R., Klineberg, E., Gupta, M., Deviren, V., Mundis, G., Hart, R., Bess, S., Lafage, V., & International Spine Study Group. (2013). The SRS-Schwab Adult Spinal Deformity Classification : Assessment and Clinical Correlations Based on a Prospective Operative and Nonoperative Cohort. *Neurosurgery*, *73*(4), 559-568. https://doi.org/10.1227/NEU.0000000000000012

Udupa, J. K., Leblanc, V. R., Zhuge, Y., Imielinska, C., Schmidt, H., Currie, L. M., Hirsch, B. E., & Woodburn, J. (2006). A framework for evaluating image segmentation algorithms. *Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society*, *30*(2), 75-87. https://doi.org/10.1016/j.compmedimag.2005.12.001

Wang, L., Xu, Q., Leung, S., Chung, J., Chen, B., & Li, S. (2019). Accurate automated Cobb angles estimation using multi-view extrapolation net. *Medical Image Analysis*, *58*, 101542. https://doi.org/10.1016/j.media.2019.101542

Warfield, S. K., Zou, K. H., & Wells, W. M. (2004). Simultaneous truth and performance level estimation (STAPLE) : An algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, *23*(7), 903-921. https://doi.org/10.1109/TMI.2004.828354

Yang, Y., Tang, Y., Gao, R., Bao, S., Huo, Y., McKenna, M. T., Savona, M. R., Abramson, R. G., & Landman, B. A. (2021). Validation and estimation of spleen volume via computer-assisted segmentation on clinically acquired CT scans. *Journal of Medical Imaging (Bellingham, Wash.)*, *8*(1), 014004. https://doi.org/10.1117/1.JMI.8.1.014004

Yeung, K. H., Man, G. C. W., Lam, T. P., Ng, B. K. W., Cheng, J. C. Y., & Chu, W. C. W. (2020). Accuracy on the preoperative assessment of patients with adolescent idiopathic scoliosis using biplanar low-dose stereoradiography : A comparison with computed tomography. *BMC Musculoskeletal Disorders*, *21*(1), 558. https://doi.org/10.1186/s12891-020-03561-2