Automatisation du traitement des imageries tridimensionnelles dento-maxillo-faciales par apprentissage profond : application à la segmentation et à la céphalométrie

Gauthier Dot

HESAM UNIVERSITÉ

Arts et Métiers Sciences et Technologies

ÉCOLE DOCTORALE SCIENCES DES MÉTIERS DE L'INGÉNIEUR
Institut de Biomécanique Humaine Georges Charpak – Campus de Paris

THÈSE

présentée par : **Gauthier DOT**

soutenue le : **4 juillet 2022**

pour obtenir le grade de : **Docteur d'HESAM Université**

préparée à : **École Nationale Supérieure d'Arts et Métiers**

Spécialité : **Biomécanique**

Automatisation du traitement des imageries tridimensionnelles dento-maxillo-faciales par apprentissage profond : application à la segmentation et à la céphalométrie

THÈSE dirigée par :
M. Thomas SCHOUMAN et M. Philippe ROUCH

et co-encadrée par :
M. Laurent GAJNY

Jury

Mme Marie-José BOILEAU, Professeur des Universités - Praticien Hospitalier, Université de Bordeaux	Présidente
Mme Elsa ANGELINI, Maître de Conférences HDR, LTCI, Télécom ParisTech, Université Paris Saclay, Paris, France	Rapporteure
M. Thomas COLARD, Professeur des Universités - Praticien Hospitalier, Université de Lille	Rapporteur
Mme Ninon BURGOS, Chargée de recherche CNRS, ARAMIS Lab, ICM	Examinatrice
M. Yannick TILLIER, Professeur des Universités, CEMEF, MINES ParisTech	Examinateur
M. Thomas SCHOUMAN, Maître de Conférences des Universités – Praticien Hospitalier, Sorbonne Université	Examinateur
M. Philippe ROUCH, Professeur des Universités, IBHGC, Arts-et-Métiers Paris	Examinateur
M. Laurent GAJNY, Maître de Conférences, IBHGC, Arts-et-Métiers Paris	Examinateur
M. Guillaume DUBOIS, Directeur général Materialise France et Professeur associé à temps partiel, IBHGC, Arts-et-Métiers Paris	Invité

T H È S E

Remerciements

Ce travail de thèse n'aurait été possible sans l'implication d'un grand nombre de personnes que je tiens à remercier ici.

Je remercie en premier lieu mon équipe d'encadrement, pour m'avoir fait confiance dans le choix de cette thématique de recherche et m'avoir guidé dans ce travail. Leur complémentarité a été cruciale pour mener à bien ce projet. Merci à Thomas Schouman pour son expertise, ses conseils avisés et son accompagnement sur les aspects cliniques et réglementaires de cette recherche. Merci à Laurent Gajny pour son investissement, sa disponibilité, son expérience et son soutien qui ont été essentiels pour mener à bien cette thèse. Merci à Philippe Rouch pour sa confiance, ses encouragements et nos discussions m'ayant encouragé à prendre du recul sur ce travail.

Je tiens à remercier les membres du jury qui ont accepté de prendre le temps de juger et d'améliorer ce travail. Je remercie particulièrement Marie-José Boileau, qui m'a fait l'honneur de présider le jury, ainsi qu'Elsa Angelini et Thomas Colard pour leur investissement en tant que rapporteurs. Un grand merci à Ninon Burgos et Yannick Tillier pour leur participation en tant qu'examinateurs. Merci à Guillaume Dubois pour son soutien et sa participation au jury.

Je remercie grandement la Fondation des Gueules Cassées pour l'attribution d'une bourse de recherche et la Société Française d'Orthopédie Dento-Faciale pour la participation au financement de ce projet de recherche. Merci à la société Materialise pour sa contribution à l'évaluation de nos résultats.

Je remercie chaleureusement l'ensemble des membres de l'Institut de Biomécanique Humaine Georges Charpak pour leur accueil et leur accompagnement. Merci à Sébastien Laporte qui a la difficile mission de diriger ce laboratoire. Merci à mes co-doctorants qui ont activement participé à rendre ces années très agréables malgré les distanciations qui nous ont été imposées. Merci à Sylvain Persohn et Marine Souq, sans qui le laboratoire ne serait pas ce qu'il est.

Je remercie les consœurs et confrères hospitalo-universitaires qui m'ont donné l'envie de suivre leurs traces et m'ont encouragé dans cette démarche. Merci à Frédéric Rafflenbeul pour son implication active dans ce travail.

Merci à ma famille et mes amis pour leur soutien et leur présence.
Merci Anastasia, pour tout ☺.

Table des matières

Introduction

L'analyse céphalométrique est un outil de diagnostic et de planification thérapeutique utilisé en routine clinique par les orthodontistes. Elle repose sur la localisation de points céphalométriques sur des imageries dento-maxillo-faciales, ces points étant par la suite utilisés pour mesurer des distances et des angles. L'analyse est classiquement effectuée sur une radiographie bidimensionnelle (2D) de profil ou de face d'un patient. Si cette approche reste suffisante pour la plupart des patients, elle ne permet pas d'analyser finement les structures superposées ou les symétries dento-maxillo-faciales. Ainsi, l'utilisation d'imageries tridimensionnelles (3D) permettrait d'améliorer le diagnostic et la planification des traitements pour certains patients présentant des anomalies dento-maxillo-faciales complexes (fentes labio-palatines, syndromes cranio-faciaux, traitements orthodontico-chirurgicaux, etc.).

Après l'acquisition d'une imagerie 3D, la réalisation d'une analyse céphalométrique 3D repose sur trois grandes étapes : (1) la reconstruction des modèles surfaciques 3D des zones d'intérêt, processus appelé « segmentation » ; (2) le placement des points céphalométriques ; (3) la réalisation des mesures. Ce traitement repose actuellement sur de nombreuses étapes de traitement manuel, nécessitant plusieurs niveaux de validation, du temps et des opérateurs formés. Cette complexité de mise en œuvre est un frein à l'utilisation clinique de l'analyse céphalométrique 3D.

L'automatisation du traitement des images par les ordinateurs a été révolutionné en 2012 par le développement de l'apprentissage profond supervisé, domaine de l'intelligence artificielle utilisant des réseaux de neurones. Cette technologie repose sur l'entraînement de modèles mathématiques à partir de bases de données comprenant de nombreuses images annotées par des opérateurs. Une fois un modèle entraîné, ce dernier peut effectuer des prédictions sur des images jamais vues auparavant. Dans le domaine de la santé, nous assistons actuellement à une croissance exponentielle des applications de l'apprentissage profond supervisé. En odontologie, de nombreux auteurs ont déjà proposé l'utilisation de cet outil pour automatiser des tâches comme la détection de lésions carieuses sur des radiographies ou des photographies intra-buccales. Ces premières applications sont très prometteuses, mais les publications souffrent de limites méthodologiques (bases de données réduites et/ou peu représentatives, critères d'évaluation ayant peu d'intérêt clinique, etc.) qui ne permettent pas encore d'assurer l'applicabilité clinique de leurs résultats.

Concernant l'analyse céphalométrique 3D, l'utilisation de modèles d'apprentissage profond a été proposé pour l'automatisation de la segmentation et du placement des points céphalométriques, avec de meilleurs résultats que les méthodes automatiques ou semi-automatiques utilisées jusqu'alors. Ces résultats très prometteurs, souvent proches des opérateurs humains, peuvent faire espérer une utilisation clinique de ces outils. Cependant, les études publiées jusqu'à maintenant restent préliminaires et nous ne savons pas comment les algorithmes se comporteraient dans le cadre d'une utilisation en pratique clinique. Une limite majeure provient des bases de données utilisées pour entraîner et tester les modèles, qui ne sont pas assez vastes et diversifiées pour assurer un comportement adéquat des outils face à de nouvelles imageries. Une autre limite provient des méthodes utilisées pour construire les références utilisées pour entraîner les modèles, dont l'élaboration n'est pas souvent décrite alors que ce sont elles qui conditionnent ce que le modèle « apprendra ». Enfin, les méthodes d'évaluation de ces algorithmes sont souvent basées sur des critères provenant des disciplines techniques et n'ont donc pas toujours d'applicabilité clinique.

L'objectif principal de notre travail a été de mettre en œuvre des modèles d'apprentissage profond supervisé permettant d'effectuer de façon automatisée la segmentation et le placement de points céphalométriques sur des imageries 3D dento-maxillo-faciales. La validation de ces modèles a été effectuée sur une base de données originale de patients présentant des malformations faciales variées et marquées, en comparant la performance de l'algorithme avec celle d'experts sur la base de critères présentant une pertinence clinique.

La première partie de ce manuscrit est consacrée à une description du contexte de l'analyse céphalométrique 3D et de l'automatisation par apprentissage profond. Notre revue systématique de la littérature, portant sur la fiabilité du placement automatisé des points céphalométriques 3D, permet de faire le point sur l'état actuel de la recherche dans le domaine. En particulier, nous mettons en avant les limites retrouvées dans les publications actuelles et les recommandations de bonnes pratiques visant à diminuer les biais des publications futures. Ce travail a été publié en 2020.

La deuxième partie de notre travail concerne l'automatisation de la segmentation des imageries dento-maxillo-faciales 3D via un modèle d'apprentissage profond supervisé. Notre étude présente l'entraînement d'un modèle dédié à la segmentation des tissus d'intérêt pour l'orthodontie et la chirurgie maxillo-faciale. Ce modèle permet l'obtention de résultats cliniquement viables sur un jeu de données de 153 scanners, réduisant très fortement la charge de travail manuelle nécessaire à la segmentation. Cette étude a été publiée en 2022. Afin d'évaluer le comportement de cet algorithme

vis-à-vis de nouvelles données externes, nous présentons des résultats complémentaires issus d'une nouvelle base de données de 25 imageries.

La troisième partie se concentre sur la céphalométrie tridimensionnelle. Une première étude s'intéresse à la fiabilité du placement manuel des points céphalométriques 3D. Trois opérateurs ont localisé deux fois une série de points céphalométriques afin d'évaluer la répétabilité et la reproductibilité de cette tâche manuelle, considérée aujourd'hui comme la référence clinique. Ce travail, publié en 2021, nous a permis d'évaluer la « référence » à atteindre dans le cadre d'une deuxième étude s'intéressant à l'automatisation du placement des points céphalométriques 3D via des modèles d'apprentissage profond. Testée sur un jeu de données de 38 scanners, notre méthode nécessite encore des améliorations pour localiser les points dentaires mais elle s'est montrée aussi efficace que les experts cliniciens pour l'évaluation des indices céphalométriques squelettiques. Ce travail a été publié en 2022.

Enfin, nous illustrons et discutons dans la quatrième et dernière partie le potentiel clinique de l'ensemble de nos résultats, dans le contexte de la planification chirurgicale maxillo-faciale, à travers trois cas cliniques.

PARTIE A :

Contexte et revue de la littérature

1 : L'analyse céphalométrique

1.1. Principe et utilisation clinique de l'analyse céphalométrique

1.1.1. Une analyse classiquement effectuée en 2D

L'analyse céphalométrique (ou céphalométrie) est une méthode diagnostique standardisée utilisée quotidiennement par les orthodontistes. Elle repose sur des mesures linéaires et angulaires effectuées sur des imageries radiographiques. A l'heure actuelle, la méthode de référence pour effectuer cette procédure est la localisation manuelle de points d'intérêts anatomiques (ou points céphalométriques) permettant d'effectuer des tracés céphalométriques sur des radiographies 2D du massif dento-maxillo-facial d'un patient (Figure 1). Ces tracés permettent ensuite d'effectuer des « mesures céphalométriques » (angles, distances ou rapports de distances) qui sont comparées à des « normes céphalométriques » permettant de poser un diagnostic et d'établir des objectifs de traitement (Proffit et al. 2018).

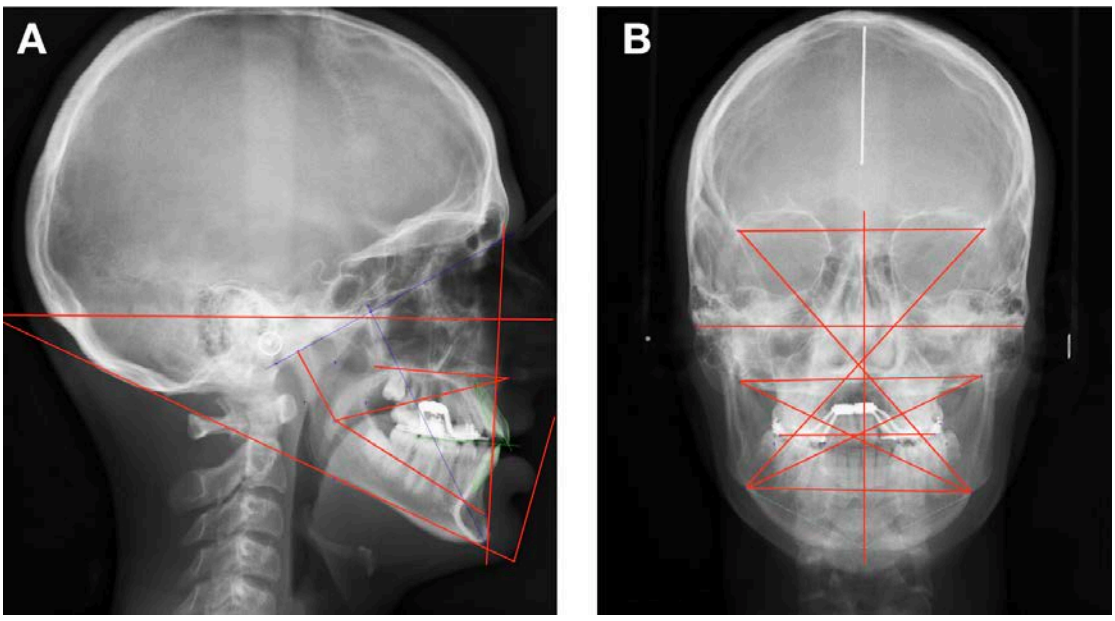


Figure 1 : Exemple de tracés céphalométriques réalisés sur des radiographies 2D : A. Radiographie de profil ; B. Radiographie de face.

De nombreuses analyses ont été proposées, se basant sur différentes approches de traitement et reposant sur des points, mesures, ou normes céphalométriques variables. Les praticiens choisissent généralement d'utiliser une analyse parmi les autres en fonction de leurs affinités et habitudes

cliniques. Afin d'illustrer la variabilité des approches, nous pouvons citer deux analyses largement utilisées en France :

- l'analyse de Tweed repose des mesures angulaires qui sont comparées à des « normes » établies à partir des mesures moyennes d'un échantillon de sujets caucasiens considérés comme « normaux ». Par exemple, les angles SNA (Sella-Nasion-Point A) et SNB (Sella-Nasion-Point B) permettent d'évaluer les positions antéro-postérieures respectives du maxillaire et de la mandibule ; l'angle FMA (entre le plan de Francfort et le plan basilaire de la mandibule) permet d'évaluer la typologie verticale du patient (Figure 2A) ;
- l'analyse de Delaire est dite architecturale, cherchant à comparer les proportions des différentes structures du massif cranio-facial d'un individu les unes par rapport aux autres. Les lignes et angles crâniens et faciaux construits à partir des points céphalométriques sont associés à l'âge, au sexe et à la typologie pour permettre de caractériser la position optimale du schéma facial du patient (Figure 2B).

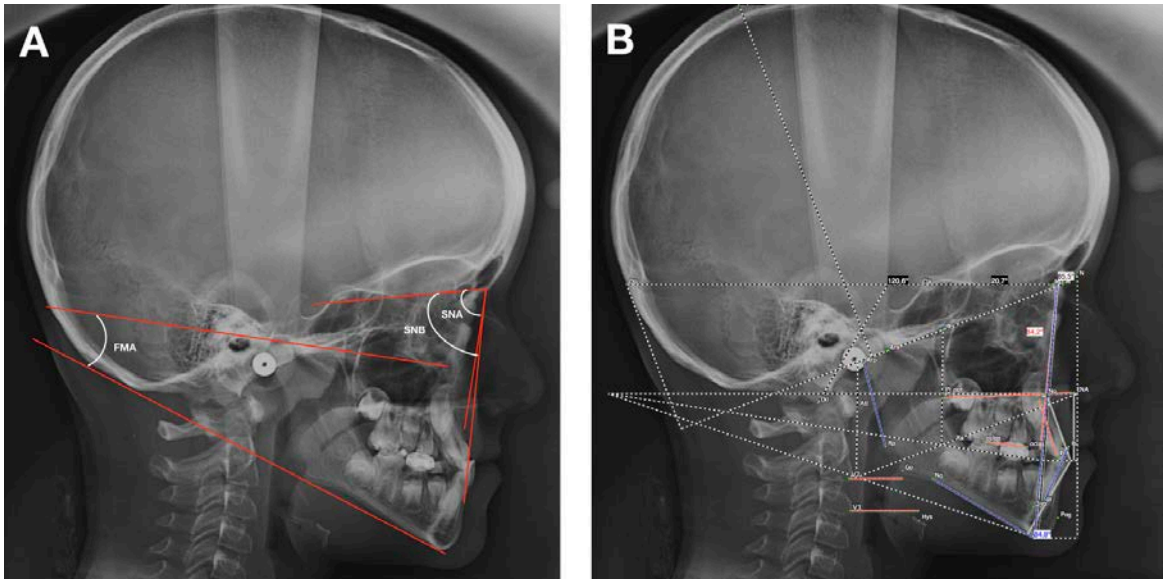


Figure 2 : Exemples d'analyses céphalométriques 2D, chez un même patient : A. Analyse partielle de Tweed, présentant les angles SNA, SNB et FMA ; B. Analyse de Delaire.

Si le placement manuel des points céphalométriques reste la référence, les travaux de recherche concernant le placement automatisé de ces points se sont fortement développés ces dernières années (Wang et al. 2016). Ces approches pourraient être aussi fiables que des cliniciens expérimentés, mais leurs méthodes de validation restent à améliorer (Schwendicke, Chaurasia, et al. 2021). Certaines de ces approches automatisées sont déjà commercialisées, comme CellmatIQ (Hambourg, Allemagne), ORCA AI (Herzliya, Israel) ou WebCeph (Gyeonggi-do, Corée).

L'utilisation de radiographies 2D présente des limites inhérentes à la technique, comme une superposition des structures bilatérales ou une distorsion des images avec certaines zones agrandies et d'autres diminuées (Gribel et al. 2011). L'imagerie 2D se montre donc peu adaptée pour le diagnostic de certains patients qui présentent des malformations faciales (dysmorphies faciales) marquées comme des asymétries importantes ou des syndromes cranio-faciaux.

1.1.2. L'analyse céphalométrique 3D

Pour ces patients dits « complexes », il a été montré que des imageries 3D pouvaient améliorer le diagnostic et la planification des traitements (American Academy of Oral and Maxillofacial Radiology 2013; Kapila and Nervina 2015). Afin d'exploiter au maximum ces données, il a été proposé d'effectuer des analyses céphalométriques sur ces imageries 3D.

A l'heure actuelle, il n'existe pas d'analyse céphalométrique 3D validée et largement utilisée cliniquement. Cela pourrait s'expliquer par deux freins principaux :

- le premier est la complexité de la visualisation des imageries 3D et du placement des points céphalométriques 3D, sujet qui constituera le cœur de notre travail ;
- le second est la complexité de l'interprétation des analyses, du fait de l'habitude des cliniciens pour les analyses 2D et de la quantité d'informations décuplée avec l'ajout de la troisième dimension. Cette question reste un sujet actif de recherche et dépasse le cadre de notre travail, mais nous présenterons ici brièvement différentes méthodes proposées dans la littérature.

Swennen et al. ont proposé en 2006 un atlas et un manuel de céphalométrie 3D, détaillant l'adaptation 3D de nombreux points céphalométriques traditionnellement utilisés dans les analyses 2D (Swennen, Schutyser, and Hausamen 2006). Certains de ces points sont utilisés pour construire des plans céphalométriques de référence (Swennen, Schutyser, Barth, et al. 2006). A partir de ces points, de nombreuses mesures peuvent être effectuées :

- mesures linéaires projetées (distance entre la projection de deux points sur un des plans de référence) ;
- mesures linéaires 3D (distance euclidienne entre deux points) ;
- mesures angulaires projetées (angle entre trois points projetés, ou deux plans projetés, ou deux points et un plan projetés) ;
- mesures orthogonales (distance entre un point et sa projection sur un des plans de référence) (Figure 3) ;

- mesures de proportions (ratio entre deux mesures).

Afin d'en déduire un diagnostic clinique, ces mesures doivent ensuite être comparées aux normes qui ont été établies sur des imageries 2D. Pour certains points, il est également possible de comparer les valeurs de l'hémiface droite et gauche, ce qui permet par exemple d'évaluer la symétrie à partir de mesures orthogonales (Figure 3).

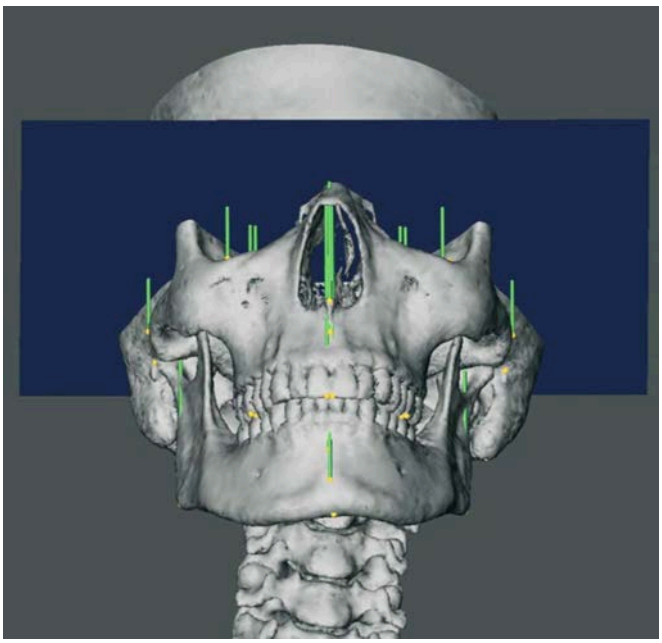


Figure 3 : Exemple de mesures orthogonales par rapport à un plan horizontal. Source : (Swennen, Schutyser, and Hausamen 2006).

Gateno et al. ont proposé en 2011 une analyse céphalométrique 3D à visée de planification chirurgicale, principalement basée sur des mesures 2D des projections des points 3D (Gateno et al. 2011). Ces mesures permettent d'utiliser les normes céphalométriques établies précédemment pour les analyses céphalométriques 2D. Cependant, les auteurs détaillent les limites géométriques de ce type de mesures chez des sujets présentant des asymétries : la projection des points sur un repère facial de référence peut fausser les mesures angulaires (Figure 4). La solution proposée est de construire un « repère monde » (massif facial) et deux « repères locaux » (maxillaire et mandibule) à partir des points céphalométriques. En fonction de leur localisation anatomique, les projections des points se font sur ces différents repères.

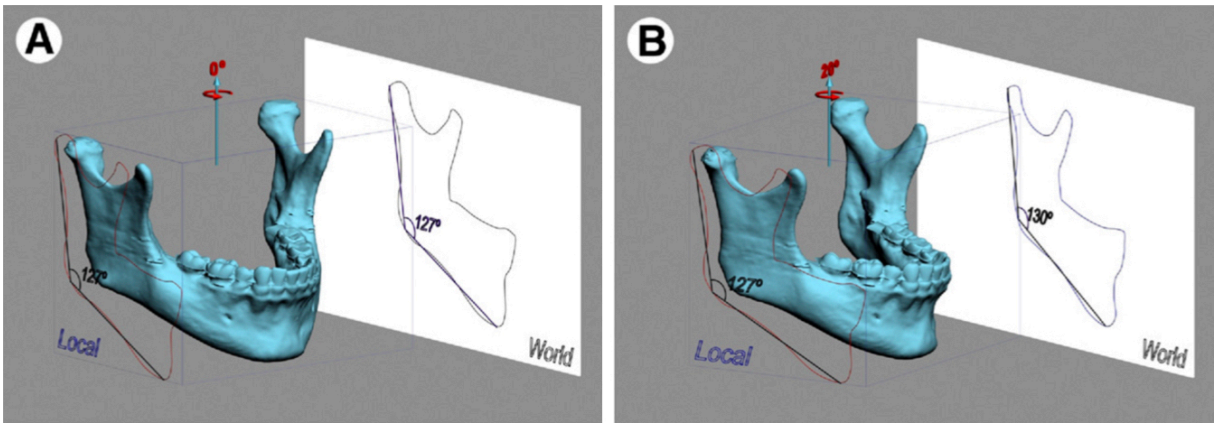


Figure 4 : Illustration de l'impact de la rotation de la mandibule sur la mesure de l'angle goniaque dans le « repère monde » : A. Mandibule centrée, l'angle goniaque est le même dans le repère local mandibulaire et le repère monde ; B. Mandibule en rotation, la projection de l'angle goniaque est augmentée dans le repère monde. Source : (Gateno et al. 2011).

A partir de la construction de deux plans 3D, trois mesures angulaires 2D peuvent être calculées selon les trois axes principaux du déplacement : le roulis (*« roll »*), le lacet (*« yaw »)* et le tangage (*« pitch »*) (Figure 5). Yatabe et al. ont proposé des interprétations cliniques de ces mesures : par exemple dans le cas de l'angle FMA, le tangage correspondant à la mesure du FMA classiquement effectuée sur une téléradiographie de profil, alors que roulis correspond à une mesure pouvant être effectuée sur une radiographie de face (Yatabe et al. 2019).

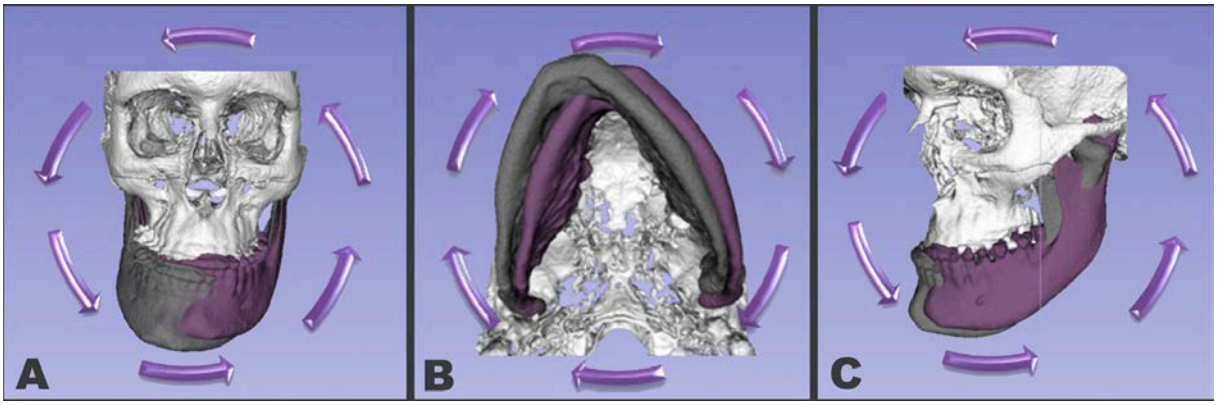


Figure 5 : Illustration schématique des trois composantes permettant de décomposer des angles 3D : A. Roulis (*« roll »*) ; B. Lacet (*« yaw »)* ; C. tangage (*« pitch »*). Source : (Yatabe et al. 2019).

Lee et al. ont proposé en 2014 une adaptation tridimensionnelle de l'analyse céphalométrique architecturale de Delaire (Lee et al. 2014). Les auteurs proposent une adaptation 3D de la description de chacun des points historiques 2D, afin de permettre la construction des plans de référence (Figure 6). Cependant, chez un même sujet, les mesures obtenues avec l'analyse 2D et l'analyse 3D proposée

ne sont pas directement comparables. Les auteurs recommandent donc d'établir de nouvelles normes et une nouvelle approche spécifique à l'analyse 3D.

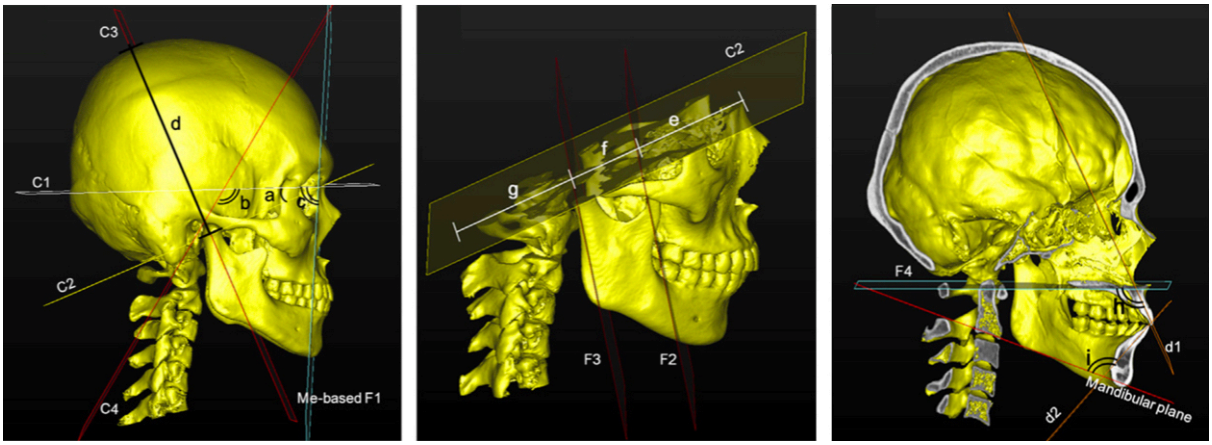


Figure 6 : Adaptation 3D de l'analyse céphalométrique de Delaire. Source : (Lee et al. 2014).

Enfin, Treil et al. ont proposé dès l'année 2000 en France une analyse 3D reposant sur 14 points osseux (7 paires de structures anatomiques) délimitant la « charpente maxillo-faciale » (Figure 7) (Treil et al. 2020). Ces points sont originaux et ne peuvent être localisés sur des imageries 2D car ils sont tous localisés sur des foramens du nerf trijumeau. Ce choix est justifié par les auteurs par trois éléments : (1) la localisation de ces points est reproductible ; (2) ces points sont en cohérence avec le rôle du nerf trijumeau comme matrice de l'embryogénèse maxillo-mandibulaire ; (3) ces points présentent une distribution homogène permettant de couvrir l'ensemble cranio-facial. Des normes pour les différentes mesures ont été publiées (Faure et al. 2005; Treil et al. 2020). Au niveau dentaire, les axes d'inertie des dents et groupes de dents sont calculés afin d'évaluer les symétries (Figure 7). Cette analyse est intégrée dans un logiciel (non commercialisé pour le moment) qui permet le placement manuel des points et la construction des axes d'inerties des dents. Une fois ces éléments connus, le logiciel présente un diagnostic synthétique, des tableaux récapitulant les différentes mesures céphalométriques et des schémas illustrant les asymétries éventuelles (Oueiss et al. 2020).

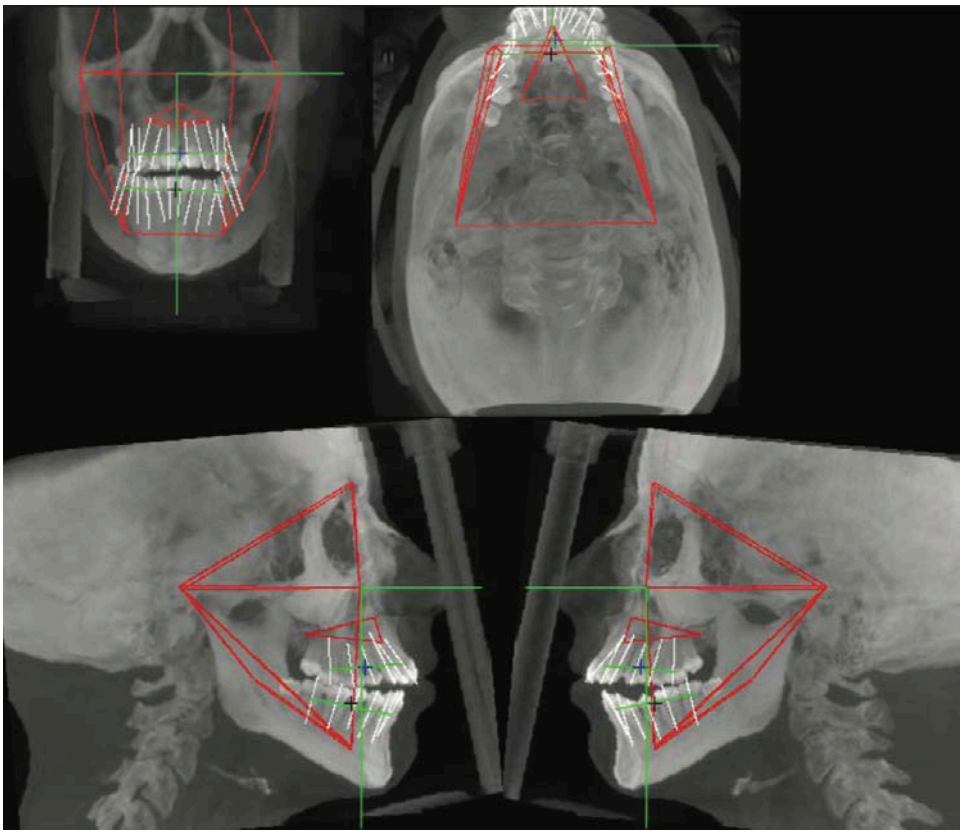


Figure 7 : Exemple d'une analyse 3D de Treil. Source : (Oueiss et al. 2020).

En l'absence de consensus sur les méthodes à adopter, nous avons fait le choix de ne pas favoriser une analyse en particulier. Ainsi, les points céphalométriques localisés dans ce travail recoupent plusieurs de ces analyses céphalométriques 3D (voir paragraphe 1.4).

1.2. Modalités d'acquisition des imageries 3D en vue d'une analyse céphalométrique

A l'heure actuelle, les imageries 3D dento-maxillo-faciales peuvent être acquises selon deux modalités principales : la tomographie volumique à faisceau conique (CBCT pour « *cone beam computed tomography* ») ou la tomodensitométrie (TDM ou CT-Scan pour « *computed tomography scan* »). Les acquisitions CBCT sont aujourd'hui majoritairement utilisées en dentisterie, en raison de leur large disponibilité, leur coût limité et leur dosimétrie généralement inférieure aux acquisitions CT-Scan (Kapila and Nervina 2015). Les acquisitions CT-Scan restent principalement utilisées pour des applications en chirurgie maxillo-faciale, telles que la réalisation des planifications chirurgicales et la réalisation de guides chirurgicaux sur-mesure (Alkhayer et al. 2020). Les principaux avantages des acquisitions CT-Scan par rapport aux acquisitions CBCT sont une meilleure visualisation des tissus mous, un meilleur comportement face aux artéfacts métalliques, un champ d'acquisition pouvant être

plus large et une calibration permettant une standardisation des niveaux de gris en fonction des tissus (unités Hounsfield) (Pauwels et al. 2015; Alkhayer et al. 2020).

Après l'acquisition, les coupes d'imagerie sont exportées au format DICOM afin d'être traitées dans des logiciels dédiés. Les mesures angulaires ou de distance réalisées sur les imageries obtenues avec les machines CBCT et CT-Scan ont été reconnues comme fiables, autorisant l'utilisation de ces acquisitions dans le cadre d'analyses céphalométriques tridimensionnelles (Smektała et al. 2014).

Les acquisitions CBCT et CT-Scan étant des imageries ionisantes, un strict respect des indications de leur utilisation est nécessaire. La réalisation d'une analyse céphalométrique nécessite un champ d'acquisition étendu à l'ensemble du massif facial, ce qui correspond actuellement à des indications très spécifiques de l'utilisation des imageries 3D : patients présentant des fentes alvéolo-palatines ou des dysmorphies faciales majeures et/ou patients candidats à une chirurgie orthognathique (Figure 8) (American Academy of Oral and Maxillofacial Radiology 2013; Kapila and Nervina 2015). Ces indications pourraient évoluer avec la commercialisation de CBCT dits ultra basse dose (« *ultra low-dose* » - ULD), dont la dose effective est inférieure à celle d'une radiographie panoramique et d'une téléradiographie de profil (van Bunningen et al. 2022). L'évaluation de la qualité diagnostique de ce type d'acquisitions ULD reste un sujet de recherche actif.

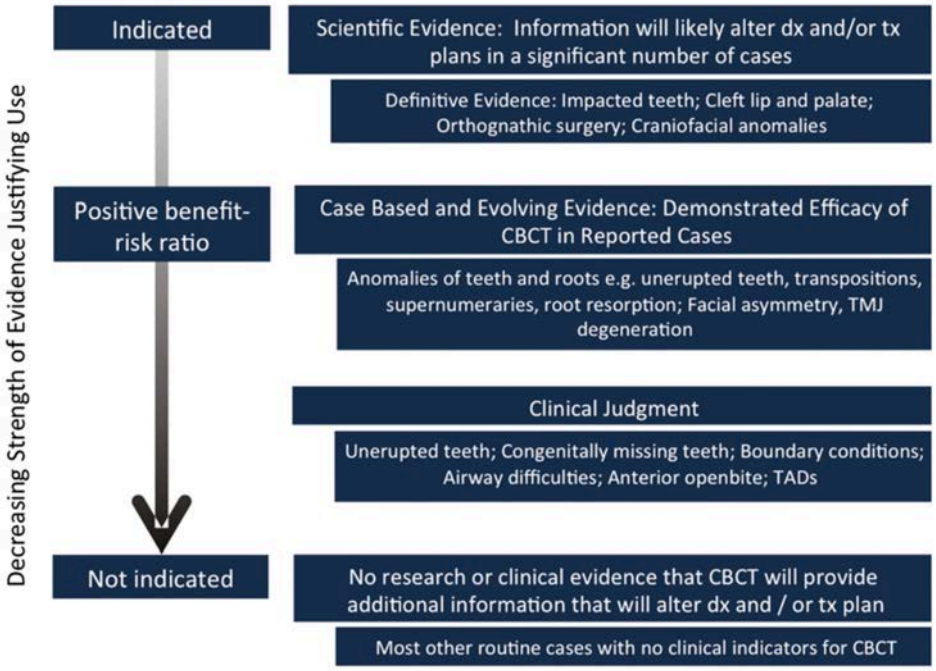


Figure 8 : Scénarios cliniques pour lesquels l'utilisation d'une imagerie 3D pourrait être indiquée, sur la base des preuves de recherche disponibles. Source : (Kapila and Nervina 2015).

Récemment, il a été proposé d'utiliser des acquisitions de type imagerie à résonnance magnétique (IRM) pour la segmentation des os du massif cranio-maxillo-facial et la réalisation de mesures céphalométriques (Zhao et al. 2018; Juerchott et al. 2020). L'avantage majeur de ces acquisitions IRM est leur absence d'irradiation ionisante. Ces résultats préliminaires très encourageants laissent entrevoir de potentielles applications cliniques à venir, qui sortent du cadre de notre travail.

1.3. Segmentation des imageries 3D dento-maxillo-faciales

1.3.1. Segmentation manuelle ou semi-automatisée

La plupart des analyses céphalométriques 3D proposées dans la littérature reposent sur des points anatomiques situés sur des surfaces osseuses. Pour effectuer les analyses et visualiser les résultats, il est nécessaire de traiter les données d'imagerie par un procédé de segmentation. Ce procédé consiste à sélectionner (« colorier ») les zones d'intérêt sur chacune des coupes d'imagerie (Figure 9).

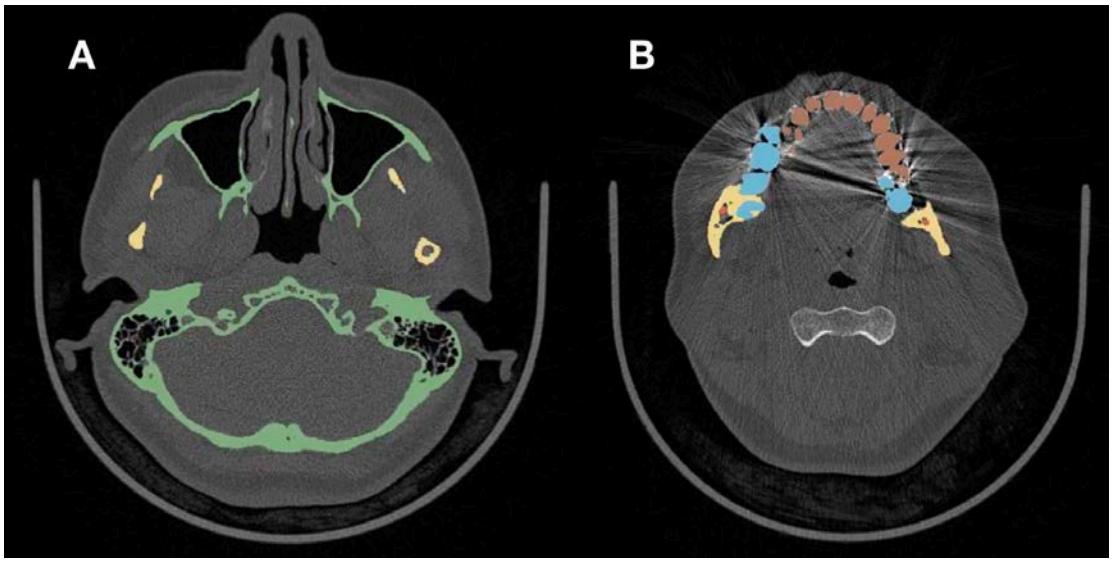


Figure 9 : Exemple de segmentation d'une acquisition CT-Scan (les zones colorées sont segmentées) : A. Coupe axiale passant par les sinus maxillaires et la branche mandibulaire ; B. Coupe axiale passant par les dents maxillaires et mandibulaires.

A partir de la segmentation, un maillage 3D de la surface externe de chacun des tissus segmentés peut être créé. Ce maillage permet d'obtenir un modèle surfacique 3D correspondant aux zones anatomiques d'intérêt. Ainsi, pour des applications orthodontiques ou maxillo-faciales nous chercherons généralement à segmenter séparément le massif facial moyen et supérieur, la mandibule, les canaux mandibulaires et les dents maxillaires et mandibulaires (Figure 10). Ces modèles surfaciques

peuvent par la suite être exportés au format stl (*STereoLithography*) pour être ouverts dans des logiciels dédiés ou imprimés en 3D.

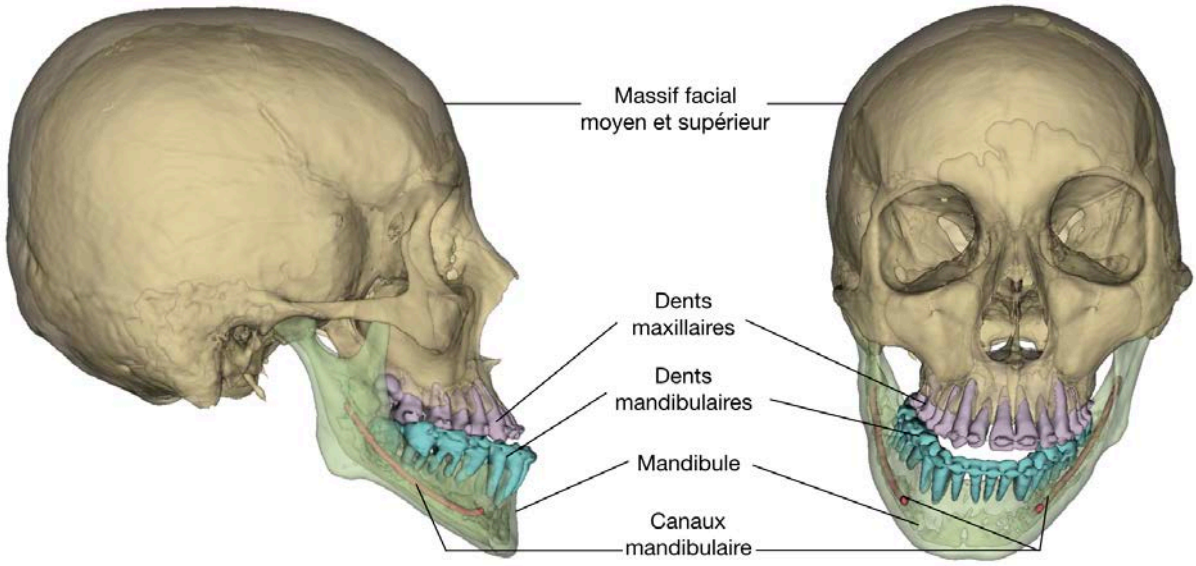


Figure 10 : Exemple de modèle surfacique 3D obtenu après segmentation d'une acquisition CT-Scan, avec les tissus d'intérêt pour une application orthodontique ou maxillo-faciale.

Différents éléments influenceront la difficulté de la tâche de segmentation :

- les contacts entre les tissus à segmenter : il est difficile de séparer les différentes zones anatomiques au niveau de leurs surfaces de contact. Concernant le squelette facial, ces contact sont souvent présents entre les dents maxillaires et mandibulaires ou entre le condyle mandibulaire et la fosse mandibulaire de l'os temporal (Torosdagli et al. 2017) ;
- les tissus à segmenter : il est par exemple plus difficile de segmenter des dents par rapport au reste des mâchoire que de segmenter l'ensemble dent et mâchoire (Wang et al. 2021) ;
- l'indication : on ne cherchera pas la même précision de reconstruction si la segmentation a vocation à être uniquement utilisée pour une visualisation de l'anatomie ou pour la fabrication de guides chirurgicaux sur-mesure (Verhelst et al. 2021) ;
- les artéfacts : la présence d'artéfacts métalliques (restaurations dentaires, appareillages orthodontiques) complique le procédé (Torosdagli et al. 2017).

Le plus souvent, ce procédé de segmentation est actuellement effectué de façon manuelle ou semi-automatisée à partir de techniques de seuillage (« *thresholding* »). L'opérateur sélectionne des niveaux de gris correspondant au mieux aux différents tissus à segmenter, puis corrige manuellement les résultats (Figure 11). Cette approche est plus facile à effectuer sur des acquisitions CT-Scan que CBCT, en raison de la calibration des niveaux de gris des imageries CT-Scan (Pauwels et al. 2015).

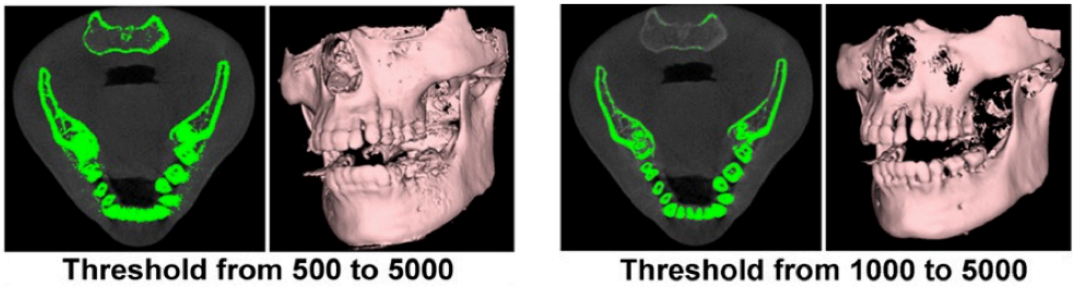


Figure 11 : Conséquence du niveau de seuillage (« *threshold* ») sélectionné pour la segmentation d'une acquisition CBCT. Source : (Pauwels et al. 2015).

Afin de comparer quantitativement le résultat de deux segmentations effectuées pour une même imagerie, la méthode d'évaluation la plus couramment utilisée est l'Indice de Sørensen-Dice (« *Dice Similarity Coefficient* » - DSC), qui mesure le chevauchement entre les volumes de prédiction et de référence (Figure 12).

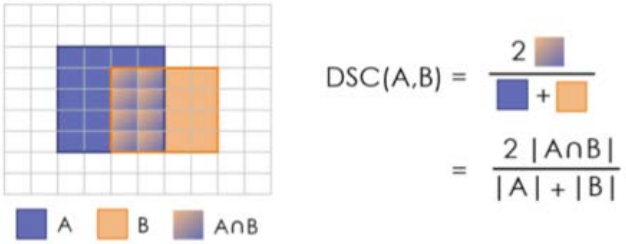


Figure 12 : Présentation visuelle et méthode de calcul du DSC. Source : (Reinke et al. 2021 Apr 13).

Les durées rapportées dans la littérature pour cette segmentation manuelle sont variables, allant de 38 minutes pour une imagerie CT-Scan sans dent et sans artéfact (Wallner, Schwaiger, et al. 2019) à plus de 5 voire 12 heures pour une acquisition CBCT (Liu et al. 2021; Wang et al. 2021). Afin d'évaluer la fiabilité de la segmentation manuelle des imageries dento-maxillo-faciales, une étude s'est intéressée à la segmentation de la mandibule sur des imageries CT-Scan sans dent et sans artéfact. Le DSC moyen entre les segmentations manuelles effectuées par 2 opérateurs sur 10 imageries était de 94.1 ± 1.2 % (Wallner, Schwaiger, et al. 2019).

Des approches semi-automatisées ont été implémentées dans différents logiciels, permettant de réduire la durée de traitement aux alentours de 10 minutes pour une imagerie CT-Scan sans dent et sans artéfact (Wallner, Schwaiger, et al. 2019). Cependant, l'expertise d'un opérateur reste nécessaire et les résultats sont inférieurs à ceux de la segmentation manuelle : la méthode présentant les meilleurs résultats dans une étude évaluant différents algorithmes semi-automatisés sur 10 imageries CT-Scan avait un DSC moyen de 85.6 ± 3.4 % (Wallner, Schwaiger, et al. 2019). Une automatisation complète de la procédure n'est pas encore disponible en routine clinique.

1.3.2. Segmentation automatisée

L'automatisation complète de la segmentation des imageries 3D dento-maxillo-faciales est un champ de recherche très actif. Afin d'évaluer quantitativement les résultats des algorithmes, il est d'usage de comparer la segmentation proposée par la méthode de référence (manuelle) avec la prédiction effectuée par la méthode automatisée, à l'aide du calcul du DSC (Figure 12). Jusqu'à l'année 2018, les méthodes les plus prometteuses reposaient sur des modèles d'atlas (Chen and Dawant 2016 Feb 2), des modèles statistiques de formes (Gollmer and Buzug 2012) ou des modèles de régressions basés sur une forêt d'arbres décisionnels (Torosdagli et al. 2017). Ces méthodes offraient des résultats encourageants mais pouvaient être mises en difficulté par les anatomies très variables retrouvées dans les imageries dento-maxillo-faciales, leurs DSC avoisinant les 90 % et restant donc inférieurs à la fiabilité de la segmentation manuelle (Wang et al. 2016; Torosdagli et al. 2017).

Depuis 2018, les travaux reposant sur l'utilisation d'algorithmes d'apprentissage profond se sont développés et présentent les résultats les plus encourageants. Davantage d'explications sur le fonctionnement des modèles d'apprentissage profond sont présentées au chapitre 3.

Nous avons réalisé une revue de la littérature incluant 11 travaux récents de segmentation basée sur l'apprentissage profond pour des applications orthodontiques ou maxillo-faciales (Tableau 1). Les acquisitions étaient au format CBCT ou CT-Scan, avec des tailles de voxels généralement inférieures à 0.5*0.5*0.5 mm^3. Les travaux portaient sur la segmentation de différents tissus :

- segmentations du massif facial supérieur et de la mandibule incluant les dents (Figure 13A) (Egger et al. 2018; Qiu et al. 2019; Torosdagli et al. 2019; Zhang et al. 2020; Liu et al. 2021; Qiu et al. 2021; Verhelst et al. 2021) ;
- segmentation séparée des dents et des surfaces osseuses (Figure 13B) (Wang et al. 2021) ;
- segmentation unitaire et reconnaissance des dents (Figure 13C) (Chung et al. 2020; Fontenele et al. 2022) ;
- segmentation des canaux mandibulaires (Figure 13D) (Lahoud et al. 2022).

La comparaison directe des résultats est difficile car les bases de données et les méthodes utilisées sont variables d'une étude à l'autre. Les résultats rapportés pour les méthodes basées sur l'apprentissage profond sont proches de la fiabilité de la segmentation manuelle, avec des DSC avoisinant les 95 %. Si ces différents travaux s'intéressent aux tissus d'intérêt pour une application orthodontique ou maxillo-faciale, aucun n'a segmenté séparément l'ensemble des tissus d'intérêt présentés en Figure 11.

Il est nécessaire de noter le nombre restreint d'imageries dans les jeux de données de test de ces études (souvent moins de 50 imageries), ce qui peut questionner sur le potentiel de généralisation de ces résultats dans un contexte clinique (voir chapitre 3). Plusieurs auteurs ont fait le choix d'exclure de leur base de données les sujets présentant des dysmorphies faciales importantes ou des artéfacts métalliques (implants, appareillages orthodontiques), ce qui pourrait constituer un biais de sélection important au vu des dysmorphies faciales et des artéfacts fréquemment rencontrés dans la population orthodontique. Malgré ces limites, des offres commerciales commencent à voir le jour, à l'instar du « *Virtual Patient Creator* » proposé par la société Relu (https://fr.relu.eu/).

D'autres applications de la segmentation automatisée d'imageries dento-maxillo-faciales par apprentissage profond ont été proposées ces dernières années, comme la segmentation des organes à risque pour la radiothérapie (Nikolov et al. 2021). Ces travaux reposent le plus souvent sur des imageries CT-Scan avec des voxels de taille importante (souvent supérieure à 1*1*1.5 mm^3), ce qui ne permet pas la réutilisation directe des modèles entraînés pour les applications précédemment décrites.

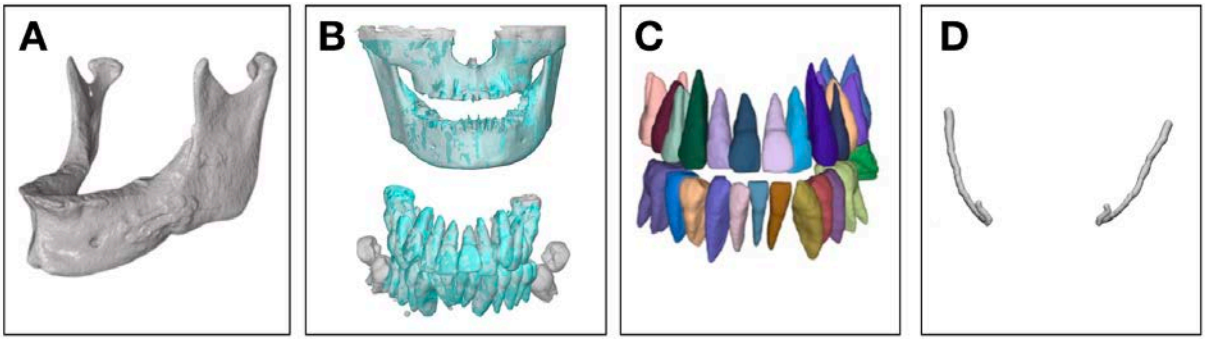


Figure 13 : Exemple des tissus segmentés dans la littérature, pour des applications orthodontiques ou maxillo-faciales : A. Segmentation de la mandibule incluant les dents ; B. Segmentation séparée des dents et surfaces osseuses ; C. Segmentation unitaire des dents ; D. Segmentation des canaux mandibulaires. Sources : (Verhelst et al. 2021; Wang et al. 2021; Chung et al. 2020; Lahoud et al. 2022).

En résumé, les modèles d'apprentissage profond de segmentation des imageries 3D pour des applications orthodontiques ou maxillo-faciales sont très prometteurs. Cependant, aucun travail n'a cherché à segmenter séparément l'ensemble des tissus qui nous intéressent : le crâne avec le massif maxillaire, la mandibule, les canaux mandibulaires et les dents maxillaires et mandibulaires (Figure 10). De plus, l'applicabilité clinique des travaux publiés reste à démontrer du fait de biais potentiels dans la sélection des imageries et du faible nombre d'imageries dans les bases de données de test.

Tableau 1 : Revue de la littérature sur la segmentation d'imageries 3D par apprentissage profond pour des applications orthodontiques ou maxillo-faciales

Étude	Base de données				Modèle	Tissus segmentées et Résultats principaux
	Sélection	Entraînement	Validation	Test		
(Egger et al. 2018)	Sélection rétrospective. Uniquement patients sans dents	10 CT		0	VGG-16 puis FCN-32s	Mandibule sans dents, DSC = 89.64 ± 1.69
(Torosdagli et al. 2019)	Sélection rétrospective. Inclus des patients syndromiques	50 CBCT (5-Fold cross-validation)		0	2D « Tiramisu » DenseNet	Mandibule & dents, DSC = 93.82
(Qiu et al. 2019)	Sélection rétrospective	52 CT	8 CT	49 CT	2D U-Net	Mandibule & dents, DSC = 88.1
(Zhang et al. 2020)	Sélection rétrospective. Patients non syndromiques	30 CT & 77 CBCT (idem que jeu d'entraînement, 5-Fold cross-validation)		0	3D U-Net customisé	Maxillaire & dents, DSC = 93.19 ± 0.89 Mandibule & dents, DSC = 93.27 ± 0.97
(Chung et al. 2020)	Sélection rétrospective. Imageries avec artéfacts métalliques	1/ 100 CBCT 2&3/ 50 CBCT		25 CBCT	1/ VGG-16 2/ R-CNN 3/ 3D U-Net	Dents individuelles, F1 score = 0.93 ± 0.03
(Qiu et al. 2021)	Sélection rétrospective. Majorité d'imageries sans artéfacts métalliques	90 CT	2 CT	17 CT	RSegCNN	Mandibule & dents, DSC = 97.48 ± 1.70
(Liu et al. 2021)	Sélection rétrospective aléatoire. Sujets avec dysmorphies (chirurgicaux)	119 CBCT & CT	17 CBCT & CT	34 CBCT & CT	3D U-Net	Maxillaire & dents, DSC = 88.5 ± 6.9 Mandibule & dents, DSC = 93.5 ± 3.4
(Wang et al. 2021)	Sélection rétrospective. Uniquement des imageries sans artéfacts métalliques	28 CBCT (4-Fold cross-validation)		0	Mixed-scale dense (MS-D) CNN	Maxillaire & mandibule, DSC = 93.4 ± 1.9 Dents, DSC = 94.5 ± 2.1
(Verhelst et al. 2021)	Sélection rétrospective et prospective d'imageries pré- et post- chirurgie orthognathique	160 CBCT		30 CBCT	3D U-Net	Mandibule & dents, DSC = 97.22 ± 0.62
(Fontenele et al. 2022)	Sélection rétrospective. Dents avec restaurations mais sans implants ou appareils orthodontiques	140 CBCT	35 CBCT	24 CBCT sans restauration 50 CBCT avec restaurations	3D U-Net	Dents unitaires sans restauration, DSC = 99 ± 2 Dents unitaires avec restauration, DSC = 97 ± 3
(Lahoud et al. 2022)	Sélection rétrospective et aléatoire	166 CBCT	39 CBCT	30 CBCT	3D U-Net	Canal mandibulaire, DSC = 77.4 ± 6.2

DSC : Dice Similarity Coefficient

1.4. Placement des points céphalométriques 3D

1.4.1. Placement manuel des points céphalométriques 3D

A l'heure actuelle, les analyses céphalométriques 3D (voir paragraphe 1.1.2) reposent sur le placement manuel des points céphalométriques. Ces points sont généralement placés sur les imageries 3D en utilisant à la fois les coupes d'imagerie et les modèles surfaciques 3D issus de la segmentation. Cette localisation manuelle est une procédure longue (environ 15 minutes) qui nécessite un opérateur formé (Hassan et al. 2013).

Un élément important qui conditionne la réussite de cette procédure manuelle est la reproductibilité intra- et inter-opérateurs du placement des points et des mesures céphalométriques. Quatre revues systématiques de la littérature sont parues ces dernières années à ce sujet (Pittayapat et al. 2014; Smektała et al. 2014; Lisboa et al. 2015; Sam et al. 2018). La dernière en date a inclus 13 études, qui portaient en majorité sur des imageries CBCT de patients ne présentant pas d'anomalies anatomiques. Quatre de ces travaux (Ludlow et al. 2009; de Oliveira et al. 2009; Frongia et al. 2012; Zamora et al. 2012) ont inclus des imageries de patients pré-chirurgicaux, évaluant la fiabilité d'adaptations 3D de points classiquement utilisés dans les analyses 2D. Ces revues systématiques rapportent que la reproductibilité du placement des points peut être considérée comme bonne (erreur moyenne <2 mm) mais certains points sont plus fiables que d'autres. En particulier, les points situés sur des zones courbes ou mal délimitées (contour de l'orbite, angle de la mandibule) sont les moins reproductibles, avec des erreurs moyennes souvent supérieures à 2 mm (Lagravère et al. 2010; Schlicher et al. 2012; Naji et al. 2014). Lagravère et al. ont proposé une méthode de classement qui a été reprise dans de nombreuses études, basée sur l'erreur moyenne de reproductibilité des points : les points présentant des erreurs inférieures à 1 mm seraient cliniquement acceptables, les points présentant des erreurs entre 1 et 2 mm seraient utiles dans la plupart des analyses et les points présentant des erreurs supérieures à 2 mm seraient à utiliser avec précaution (Figure 14) (Lagravère et al. 2010).

Peu d'études se sont intéressées à la reproductibilité des mesures céphalométriques effectuées à partir des points placés. Davantage de travaux sont nécessaires à ce sujet, puisque ce sont ces mesures qui sont utilisées pour effectuer le diagnostic clinique (Smektała et al. 2014).

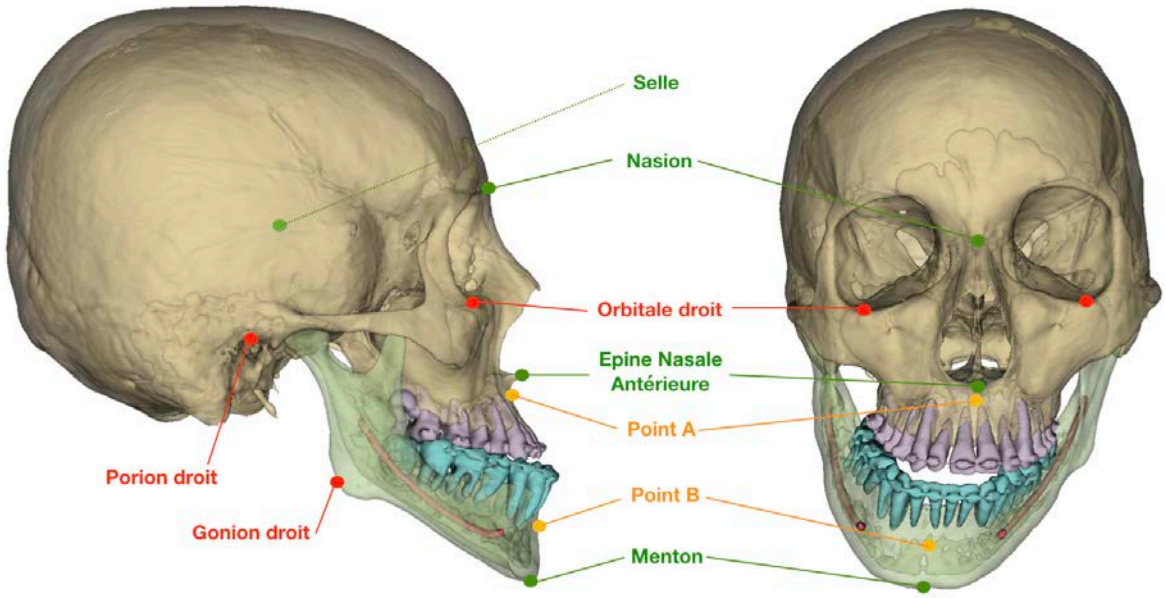


Figure 14 : Illustration de la fiabilité du placement manuel de certains points céphalométriques 3D, d'après les études parues dans la littérature. Vert : reproductibilité excellente (erreurs inférieures à 1 mm) ; Orange : reproductibilité moyenne (erreurs entre 1 et 2 mm) ; Rouge : reproductibilité faible (erreurs supérieures à 2 mm).

Les recommandations faites par les auteurs afin d'améliorer la reproductibilité du placement manuel des points sont :

- préférer des tailles de coupes fines pour les imageries (autour de 0.3 mm) (de Oliveira et al. 2009; Cattaneo et al. 2019) ;
- utiliser à la fois les coupes d'imagerie et leurs segmentations pour localiser les points d'intérêts (Ludlow et al. 2009; Hassan et al. 2013; da Neiva et al. 2015), même si cela demande plus de temps (Hassan et al. 2013) ;
- définir précisément la localisation de chaque point dans les trois plans de l'imagerie (axial, sagittal, frontal) (Ludlow et al. 2009; de Oliveira et al. 2009; Lagravère et al. 2010; Frongia et al. 2012; Hassan et al. 2013; Naji et al. 2014; Cattaneo et al. 2019).

Afin d'améliorer la reproductibilité de la localisation de ces points anatomiques, certains auteurs suggèrent de s'affranchir des points céphalométriques historiques (issus des analyses 2D) pour placer des points au niveau de structures anatomiques uniquement visualisables sur les imageries 3D. Par exemple, la localisation de points au niveau des foramens osseux est décrite comme potentiellement fiable car ces zones sont clairement délimitées (Faure et al. 2005; Lagravère et al. 2010; Schlicher et al. 2012; Naji et al. 2014; Lim et al. 2019). Cependant, un seul auteur a étudié la reproductibilité de la localisation manuelle de ces points foraminaux *in vivo*, chez des sujets non chirurgicaux (Naji et al.

2014). La difficulté qui se pose avec ces « nouveaux » points est leur utilisation dans le cadre d'une analyse céphalométrique à visée clinique : mise à part l'approche proposée par Treil et al. et présentée précédemment, la grande majorité des analyses proposées jusqu'à maintenant reposent sur des adaptations 3D des analyses historiques 2D (voir paragraphe 1.1.2). Les normes historiques 2D ne pouvant plus être utilisées, l'utilisation de « nouveaux » points est conditionnée au développement d'analyses céphalométriques et de normes spécifiques aux imageries tridimensionnelles (Faure et al. 2016; Lim et al. 2019).

Le temps et l'expertise nécessaires à la réalisation manuelle d'une analyse céphalométrique 3D la rendent peu utilisée en clinique et il n'est pas rare que des radiographies virtuelles 2D soient générées à partir d'examens d'imagerie 3D dento-maxillo-faciales pour y effectuer le placement de points céphalométriques 2D. La fiabilité des mesures effectuées sur ces images 2D a été validée, mais cela implique une perte d'information potentiellement défavorable (Moshiri et al. 2007; Kumar et al. 2008).

1.4.2. Automatisation du placement des points céphalométriques 3D

Lors du début de ce travail de recherche, aucune revue de la littérature n'avait été effectuée sur le sujet de l'automatisation de la localisation des points céphalométriques 3D. Nous avons donc réalisé une revue systématique de la littérature, portant sur les études évaluant un algorithme de placement automatique des points céphalométriques osseux sur des imageries 3D (CT-Scan ou CBCT) publiées avant le 14 mars 2019. Cette revue a été publiée dans le *International Journal of Oral and Maxillofacial Surgery* en 2020, et constitue le Chapitre 2 de ce manuscrit.

En suivant les recommandations PRISMA-DTA (Salameh et al. 2020), 11 études publiées entre 2014 et 2019 ont été incluses dans notre analyse qualitative de la littérature. Les méthodes utilisées dans les études afin de placer les points céphalométriques pouvaient être classées en 3 catégories :

- méthodes basées sur la connaissance à priori, reposant sur des définitions mathématiques des points ;
- méthodes basées sur un atlas, reposant sur un ou plusieurs modèles qui sont déformés afin de correspondre au mieux à l'image cible ;
- méthodes basées sur l'apprentissage, reposant sur une base de données annotées de référence qui est utilisée pour créer un modèle statistique ou pour entraîner un algorithme (voir Chapitre 3).

Deux études présentaient les résultats les plus prometteurs, avec des erreurs moyennes du placement automatique des points inférieures à 2 mm (Zhang et al. 2017; Torosdagli et al. 2019). Ces deux travaux ont utilisé des méthodes basées sur l'apprentissage profond, et en particulier des réseaux convolutifs d'apprentissage profond (voir Chapitre 3). De plus, ces deux méthodes proposaient d'automatiser la segmentation et la localisation des points céphalométriques au sein du même modèle. Cependant, les résultats restent préliminaires (peu de points localisés, bases de données pas assez détaillées, pas de base de données externe pour tester les modèles) et ne peuvent donc être transposables directement en clinique. Il est difficile de savoir si ces algorithmes se comporteraient bien dans le cadre d'un plus grand nombre de points à placer, chez des patients présentant des dysmorphies faciales importantes. Les résultats de cette revue systématique de la littérature (voir Chapitre 2) nous ont encouragé à faire le choix d'utiliser des modèles d'apprentissage profond pour poursuivre nos travaux de recherche dans ce domaine.

Depuis la publication de notre revue de la littérature, nous avons retenu la publication de 7 études portant sur la localisation automatique de points céphalométriques 3D via des modèles d'apprentissage profond (Tableau 2). Le nombre de points localisés est très variable, allant de 13 à 175. La grande majorité des travaux propose une approche dite « coarse-to-fine » (« grossier à fin ») permettant d'analyser dans un premier temps l'ensemble du volume d'imagerie avec une résolution dégradée (« grossier ») avant de se focaliser sur certaines zones d'intérêt en conservant la pleine résolution (« fin »). Une étude propose d'intégrer dans le même modèle la segmentation et la localisation des points céphalométriques (Liu et al. 2021). Les résultats sont très encourageants, plusieurs études rapportant des erreurs moyennes inférieures à 2 mm.

La quantité de publications parues ces trois dernières années démontre l'intérêt, l'actualité et le dynamisme de cette thématique de recherche. Cependant, les réserves émises dans notre revue systématique de la littérature restent d'actualité. La plupart de ces travaux ne détaillent pas ou peu la sélection de leurs imageries et les méthodes suivies pour localiser manuellement les points de référence. Mis à part une étude testant son modèle sur 34 CBCT et CT-Scan (Liu et al. 2021), les travaux testent leur modèle sur moins de 10 imageries ou effectuent uniquement des validations croisées sans base de données de test. L'évaluation des résultats repose le plus souvent uniquement sur la mesure des distances euclidiennes moyennes, sans détailler les résultats point par point ni par proportion de points en dessous d'un certain seuil (2 mm, 3 mm, 4 mm, etc.) (Wang et al. 2016). Selon les recommandations d'expert pour la publication d'études portant sur l'apprentissage profond en odontologie, ces éléments constituent des risques de biais importants et pourraient influencer l'applicabilité clinique de ces modèles (voir Chapitre 3) (Schwendicke, Singh, et al. 2021). Enfin, aucune

des équipes ayant travaillé sur le sujet n'a partagé les modèles ou les bases de données utilisés, ce qui ne permet pas de reproduire les résultats présentés. A notre connaissance, il n'existe pas encore de logiciel commercial offrant la possibilité de réaliser automatiquement des analyses céphalométriques 3D.

Une revue systématique de la littérature portant sur l'utilisation des algorithmes d'apprentissage profond pour l'automatisation de la localisation des points céphalométriques 2D et 3D a été publiée très récemment (Schwendicke, Chaurasia, et al. 2021). Cette revue conclut que les méthodes automatisées pourraient présenter des performances comparables à celles des cliniciens, mais les études incluses souffrent de biais importants. Cette revue n'a inclus que 4 études portant sur les points 3D et conclue que les données actuelles sont trop faibles pour pouvoir soutenir l'utilisation clinique de ces méthodes automatisées sur les imageries 3D. Les auteurs encouragent au développement de critères permettant d'évaluer l'utilité clinique de ces méthodes automatisées.

En résumé, les études rapportant l'automatisation de la localisation de points céphalométriques 3D via des algorithmes d'apprentissage profond sont très encourageantes. Cependant, les travaux publiés présentent des biais importants et leur applicabilité clinique n'est pas démontrée. Des études complémentaires sont nécessaires, incluant des jeux de données plus détaillés, des données de référence clairement établies et des critères d'évaluation des résultats cliniquement pertinents. A l'heure actuelle et à notre connaissance, les algorithmes utilisés ne sont pas partagés et il n'existe pas de logiciel commercial permettant de réaliser automatiquement des analyses céphalométriques 3D.

Tableau 2 : Revue de la littérature des études parues après 2019 portant sur la localisation de points céphalométriques 3D par apprentissage profond

Étude	Base de données				Localisation des points de référence	Modèle	Nombre de points localisés et Résultats principaux
	Sélection	Entraînement	Validation	Test			
(Ma et al. 2020)	Sélection rétrospective. Patients chirurgicaux	58 CT		8 CT	1 opérateur	CNN customisé	13 points, erreur moyenne = 5.8 ± 1.0 mm
(Yun et al. 2020)	Sélection rétrospective	22 CT 229 sets de coordonnées 3D		4 CT	1 opérateur expérimenté	CNN customisé (“coarse-to-fine”)	93 points, erreur moyenne = 3.6 mm
(Lang et al. 2020)	Sélection rétrospective aléatoire. Sujets non syndromiques	50 CT 35 CBCT	5 CBCT	10 CBCT	2 Chirurgiens maxillo-faciaux expérimentés	1/ U-Net (“coarse-to-fine”) 2/ Graph Convolution Network	60 points, erreur moyenne = 1.47 ± 0.21 mm Permet de localiser les points absents
(Kang et al. 2021)	Sélection rétrospective. « Sujets normaux » sans dysmorphie ou malocclusion	20 CT		8 CT	Moyenne de 2 experts	Double DQN (“coarse-to-fine”)	16 points, erreur moyenne = 1.96 ± 0.78 mm
(Liu et al. 2021)	Sélection rétrospective aléatoire. Sujets avec dysmorphies (chirurgicaux)	119 CBCT & CT	17 CBCT & CT	34 CBCT & CT	Chirurgiens maxillo-faciaux expérimentés	Adaptation de 3D U-Net (“coarse-to-fine”)	66 points osseux, erreur moyenne = 3.03 ± 1.96 mm 68 points dentaires, erreur moyenne = 2.10 ± 2.89 mm 41 points tissus mous, erreur moyenne = 3.34 ± 3.20 mm
(Chen et al. 2021)	Sélection rétrospective aléatoire. Sujets avec dysmorphies	80 CBCT (4-Fold cross-validation)		0	Chirurgiens maxillo-faciaux expérimentés et « vote de consensus »	3D faster R-CNN + MS-Unet (“coarse-to-fine”)	18 points, erreur moyenne = 0.89 ± 0.64 mm
(Chen, Ma, Chen, et al. 2022)	*Non précisé*	89 CBCT (3-Fold cross-validation)		0	*Non précisé*	Structure-Aware Long Short-Term Memory Network (“coarse-to-fine”)	17 points, 1.64 ± 1.13 mm

CNN : réseau neuronal convolutif

1.5. Domaines d'applications de la céphalométrie 3D automatisée

1.5.1. Orthodontie

Comme détaillé précédemment, l'analyse céphalométrique 3D serait une adaptation logique et nécessaire de l'analyse 2D pour les patients présentant des dysmorphies faciales complexes. Cette analyse permettrait de localiser l'origine des anomalies osseuses et/ou dentaires afin d'aider à l'établissement du plan de traitement orthodontique.

Cependant, l'apport clinique de ces analyses pour des patients orthodontiques n'a pas encore été suffisamment démontré (Pittayapat et al. 2014; Kapila and Nervina 2015). Des cas cliniques ont présenté l'intérêt de ces analyses dans des situations précises, mais nous n'avons pas encore de visibilité sur le type de patients qui bénéficieraient de ces analyses (Pittayapat et al. 2014; Oueiss et al. 2020). Étant donné que les imageries 3D restent pour le moment plus irradiantes que les radiographies 2D classiquement effectuées, ce type d'analyse ne peut être recommandé en routine clinique. Des travaux sur l'utilisation d'acquisitions CBCT ultra basse dose et sur le type de patients qui bénéficieraient de ces analyses 3D restent à effectuer.

1.5.2. Chirurgie orthognathique

Les imageries 3D sont déjà largement utilisées en chirurgie maxillo-faciale, en particulier dans le contexte de la planification de la chirurgie orthognathique. Cette chirurgie vise à corriger les malformations congénitales ou acquises des mâchoires, par des gestes chirurgicaux permettant le déplacement du maxillaire et/ou de la mandibule (Figure 15). La fixation des portions osseuses déplacées se fait par des plaques d'ostéosynthèse. Il a été montré que la planification de ces chirurgies était améliorée lorsqu'elle était effectuée numériquement en 3D, en particulier pour les patients présentant des dysmorphies faciales marquées (Ho et al. 2017; Alkhayer et al. 2020). Chez ces patients, la planification numérique 3D est significativement différente de la simulation classiquement effectuée en 2D, et celle-ci permet une évaluation plus précise des corrections chirurgicales à effectuer (Ho et al. 2017). La visualisation des modèles 3D et l'analyse céphalométrique 3D sont des étapes essentielles de la planification chirurgicale numérique (Gateno et al. 2011).

Afin de transférer la planification numérique 3D au bloc opératoire, des guides chirurgicaux de positionnement et/ou des plaques d'ostéosynthèse peuvent être produits sur-mesure (Figure 15) (Schouman et al. 2015). Dans ce cadre, c'est le plus souvent un partenaire industriel qui se charge de la segmentation et du placement des points céphalométriques sur les imageries 3D. Cette approche

ne permet pas aux cliniciens (chirurgien, et orthodontiste le cas échéant) de visualiser les modèles 3D et les résultats de l'analyse céphalométrique en présence du patient, pouvant compliquer les étapes de planification de la chirurgie. De plus, la production de plaques d'ostéosynthèse sur-mesure demande des segmentations de haute qualité, qui sont pour le moment effectuées selon un procédé semi-automatique restant très dépendant d'opérateurs humains.

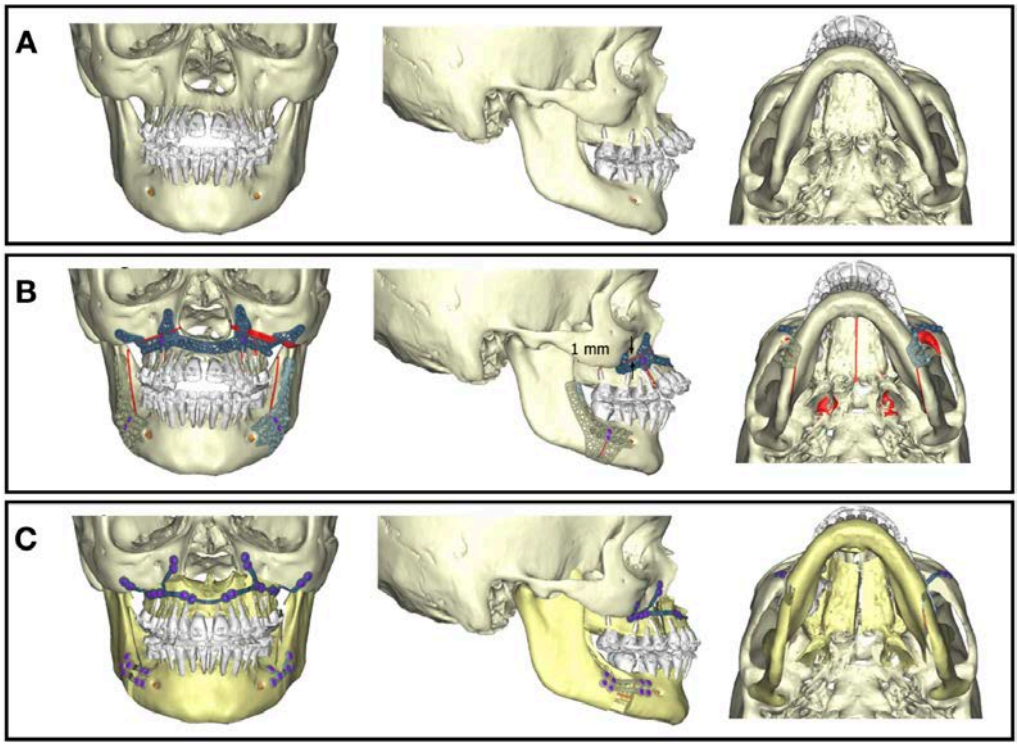


Figure 15 : Exemple de planification numérique de chirurgie orthognathique sur modèles 3D, avec déplacement maxillaire et mandibulaire : A. Anatomie initiale ; B. Guides chirurgicaux sur-mesure et traits d'ostéotomie planifiés (en rouge) ; C. Plaques d'ostéosynthèse sur-mesure.

1.5.3. Médecine légale et anthropologie médico-légale

L'anthropométrie est couramment utilisée en médecine légale, par exemple pour chercher à déterminer le sexe, l'âge biologique ou le schéma de croissance d'un individu à partir de ses imageries dento-maxillo-faciales (Thurzo et al. 2021). Ces mesures sont classiquement effectuées en 2D, mais pourraient gagner à être effectuées en 3D (Bermejo et al. 2021).

Les points céphalométriques sont également utilisés en anthropologie médico-légale, afin de chercher à déterminer l'identité d'individus dont il ne reste que le squelette (Damas et al. 2018). L'automatisation de la segmentation d'imageries 3D et de la localisation de points céphalométriques 3D pourraient également faciliter la réalisation d'analyses de morphométrie géométrique (Landi and O'Higgins 2019).

2 : Précision du placement automatisé des points céphalométriques 3D : revue systématique de la littérature

Ce chapitre a fait l'objet d'une publication dans le *International Journal of Oral and Maxillofacial Surgery* en 2020, sous le titre :

Accuracy and reliability of automatic three-dimensional cephalometric landmarking
A systematic review

Gauthier Dot[1], Frédéric Rafflenbeul[2], Marco Arbotto[1], Laurent Gajny[1], Philippe Rouch[1], Thomas Schouman[1,3]

[1] Institut de Biomécanique Humaine Georges Charpak (IBHGC), Arts et Métiers ParisTech, 151 Boulevard de l'Hôpital, 75013 Paris, France

[2] Department of Dento-Facial Orthopedics, Faculty of Dental Surgery, Strasbourg University, 8 rue Sainte-Elisabeth, 67000 Strasbourg, France

[3] Sorbonne Université, AP-HP, Hôpital Pitié-Salpêtrière, Service de Chirurgie Maxillo-Faciale, 47-83 Boulevard de l'Hôpital, 75013 Paris, France

DOI : https://doi.org/10.1016/j.ijom.2020.02.015

2.1. Abstract

The aim of this systematic review was to assess the accuracy and reliability of automatic landmarking for cephalometric analysis of 3D craniofacial images. We searched for studies that reported results of automatic landmarking and/or measurements of human head CT or CBCT scans in MEDLINE, EMBASE and Web of Science until march 2019. Two authors independently screened articles for eligibility. Risk of bias and applicability concerns for each included study were assessed using the QUADAS-2 tool. Eleven studies with test dataset sample sizes ranging from 18 to 77 images were included. They used knowledge-, atlas- or learning-based algorithms to landmark 2 to 33 points of cephalometric interest. Ten studies measured mean localization errors between manually- and automatically-detected landmarks. Depending on the studies and the landmarks, mean errors ranged from <0.50 mm to >5 mm. The two best-performing algorithms used a deep learning method and reported mean errors <2 mm for every landmark, approximating results of operator variability in manual landmarking. Risk of

bias regarding patient selection and implementation of the reference standard were found, therefore the studies might have yielded overoptimistic results. The robustness of these algorithms needs to be more thoroughly tested in challenging clinical settings. PROSPERO registration number: CRD42019119637.

2.2. Introduction

Cephalometric analysis (or cephalometry) is a standardized diagnostic and treatment evaluation method used daily by orthodontists and maxillofacial surgeons. The analysis is based on linear and angular measurements performed on radiographic images. The gold standard for this procedure is a manual detection and landmarking of meaningful anatomical structures on lateral or frontal skull radiographs called cephalograms (Leonardi et al. 2008). This X-ray technique is a two-dimensional (2D) projection of three-dimensional (3D) craniofacial structures, which leads to superimposition of bilateral structures and distortion of images, with enlargement in some areas and reduction in others (Gribel et al. 2011).

To overcome the downsides of cephalograms, several authors have proposed 3D cephalometric analysis, based on 3D craniofacial images provided by computed tomography (CT) or cone beam computed tomography (CBCT) imaging techniques (Olszewski et al. 2006; Lee et al. 2014). For now, there is no globally recognized 3D analysis or validated list of landmarks. Most of the proposed analyses have been 3D adaptations of previous 2D techniques, relying on landmarks localized on the bone surface of the skull (Olszewski et al. 2006; Lee et al. 2014). These landmarks are then used to provide cephalometric results in the form of linear (Euclidian distance between two points), angular (angle between three points or two planes) and ratio (between two linear values) measurements. An example of a set of 3D landmarks localized on a skull model is shown in Figure 16. It is suggested that 3D cephalometry could improve treatment outcomes for difficult cases (e.g. patients with craniofacial syndromes, major asymmetries/craniofacial anomalies or undergoing orthognathic surgery) when compared to traditional 2D cephalometry (Swennen, Schutyser, and Hausamen 2006; Smektała et al. 2014; Kapila and Nervina 2015). *In vitro* 3D craniofacial measurements are proven to be highly reliable, validating the possible use of CT or CBCT scans for 3D cephalometry (Smektała et al. 2014).

Manual landmarking of 3D volumes requires time and a high level of expertise and experience (Lagravère et al. 2010). Hassan et al. reported durations up to 14 minutes to place 22 landmarks (Hassan et al. 2013). Thorough training of the operators aims at reducing their identification errors in order to keep interobserver variation at a clinically acceptable level (Lagravère et al. 2010).

Reproducibility studies have shown that some landmarks are more reliable than others, with midsagittal plane landmarks usually showing greater reliability than bilateral landmarks (Sam et al. 2018). Depending on the points and the studies, inter-operator variability ranges from less than 0.5 mm to more than 2 mm (Pittayapat et al. 2014; Smektała et al. 2014; Sam et al. 2018). As a result, for the time being, this 3D technique is barely used in clinical settings and there is a lack of evidence as to which patients would benefit from it.

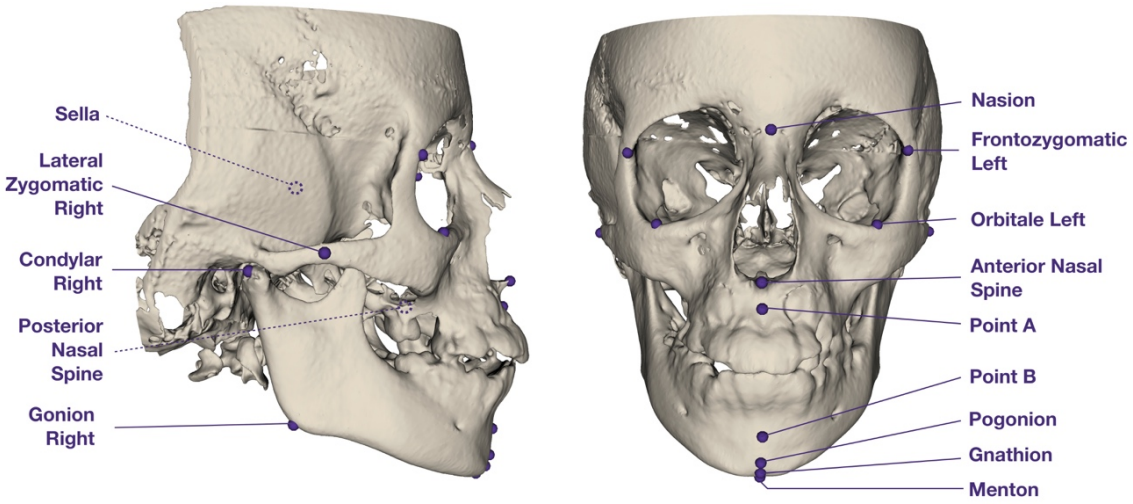


Figure 16: Example of 3D landmarks localized on a skull model, lateral right and frontal views (dotted points show approximate projections of intra-cranial landmarks).

Automatization of the 3D cephalometric landmarking process could greatly facilitate access to this diagnostic tool. It would save time and enable untrained clinicians to use 3D cephalometry on a daily basis. Automatic 3D cephalometry could be more accurate than manual landmarking by learning to average out landmarking errors (Lindner et al. 2016). Various numerical methods have been proposed, including knowledge-based, atlas-based and learning-based algorithms (Gupta et al. 2015; Zhang et al. 2017). The studies rely on different reporting methods, making it difficult to compare them. To our knowledge, neither a review of this research, nor an analytic comparison between results of different automatic 3D landmarking methods have been reported.

The aim of this systematic review is to assess the current evidence on the accuracy and reliability of automatic landmarking in comparison to manual landmarking for cephalometric analysis of 3D craniofacial images (CT or CBCT scans).

To this aim, our systematic review details the various techniques used and answers the following research questions:

- What is the accuracy of automatic 3D landmarking when compared to manual landmarking?
- How reliable are linear and angular measurements obtained through automatic landmarking when compared to manual landmarking?

2.3. Materials and Methods

2.3.1. Protocol and registration

This systematic review is reported based on the PRISMA extension for Diagnostic Test Accuracy (DTA) guidelines. In accordance with the guidelines, our protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) on the 28th of January 2019 (registration number CRD42019119637).

2.3.2. Eligibility criteria

Studies were selected according to the criteria outlined below:

- Study designs: we included in vitro and in vivo prospective and retrospective studies (clinical trials, comparative studies, validation studies or evaluation studies). We excluded book chapters, animal studies, case reports, epidemiologic studies, narrative reviews and author opinion articles.
- Population: we included studies examining the general human population, with no age limit.
- Index test: the index test of interest was automatic landmarking and/or measurements of 3D cranio-facial CT or CBCT scans. Several skeletal or dental landmarks with cephalometric interest needed to be localized in the maxillofacial area. “Automatic” meant that the landmarking or the measurements were performed by an algorithm, with minimal intervention by the operator (e.g. reorientation of the volume or manual localization of a few landmarks to run the procedure). Detailed definitions of landmarks and/or measurements needed to be provided, as well as detailed definition of the algorithm used.
- Sample: for the index test, a sample size of at least 10 images needed to be provided.
- Reference standard: manual landmarking of 3D craniofacial CT or CBCT scans.
- Timing: there was no restriction in the search period.
- Language: we included articles reported in English, French and German.

2.3.3. Information sources

Our search was performed in the following databases: MEDLINE via Pubmed, EMBASE, Web of Science and Cochrane Central Register of Controlled Clinical Trials (CENTRAL). We searched the grey literature through OpenGrey database and Google Scholar, for which we considered the first 300 results for inclusion. To ensure literature saturation, we scanned the reference lists of included studies or relevant reviews identified through the search, and handsearched for studies citing included studies. No limit has been applied as to the date of publication, and our last search was performed on the 14th of March 2019.

2.3.4. Search strategy

The publications were searched electronically by one author, using controlled index terms and relevant specific free text words. After the MEDLINE strategy was finalized, it was adapted to the syntax and subject headings of the other databases (see Supplementary Table A1 for detailed search strategy). Duplicate articles were removed after importing the lists into a reference management software (Zotero v.5.0.62).

2.3.5. Study selection

Two reviewers independently screened the resulting collection of titles and abstracts. Studies that did not pertain to the review topic were excluded. When a title or abstract was considered to be relevant by only one of the reviewers, the publication was not excluded. The full texts of the remaining publications were then retrieved and reviewers independently assessed them to decide whether these met the inclusion criteria or not. Additional information was sought from study authors where necessary to resolve questions about eligibility. Disagreements between reviewers were solved through discussion and reasons for exclusion were recorded. When the same research team had published several articles on the refinement of an algorithm, the most recent paper was included. Neither of the reviewers were blinded to the journal titles, study authors or institutions.

2.3.6. Data collection process

One author extracted data into a standardized form subsequently checked by a second author. Disagreement was resolved through discussion. Additional information was sought from study authors where necessary, but articles were excluded whenever there were three or more unanswered requests.

2.3.7. Risk of bias and applicability

To evaluate the risk of bias and applicability of each study, information was collected using a tailored checklist based on the QUADAS-2 tool (Whiting 2011) and recommendations from the Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy (Reitsma et al. 2009) (Supplementary Table A2). If there was insufficient detail reported in the study, the risk of bias was judged as "unclear". These judgements were made independently by two review authors. Disagreements were resolved first by discussion and then by consulting a third author for arbitration.

2.3.8. Diagnostic accuracy measures

The diagnostic accuracy measures reported were the mean differences and standard deviations expressed in mm (Euclidean distances), in degrees (angles) or in ratios (proportional measurements) between the automatic and the manual methods.

2.3.9. Synthesis of results

A systematic narrative and qualitative synthesis was provided with information presented in the text and tables to summarize and explain the characteristics and findings of the included studies. The narrative synthesis explored the relationship and findings both within and between the included studies. A meta-analysis was not possible because of the heterogeneity of the methodologies used in the selected studies.

2.3.10. Data availability

All the data generated or analysed during this study is included in this published article (and its Supplementary Materials).

2.4. Results

2.4.1. Study selection

The flow chart of the selection process for inclusion of articles in this study is outlined in Figure 17. A total of 654 manuscripts were selected for the screening phase (see Supplementary Table A1 for detailed results for each database) and 599 studies were excluded following abstract/title assessment. Following full-text review, a further 44 papers were excluded and the reasons for it were recorded, leaving 11 studies as eligible for inclusion in the qualitative synthesis. Among them, 10 studies (Shahidi et al. 2014; Gupta et al. 2015; Zhang et al. 2016; Codari et al. 2017; Zhang et al. 2017; de Jong et al.

2018; Montúfar et al. 2018; Neelapu et al. 2018; O'Neil et al. 2019; Torosdagli et al. 2019) were related to our research question 1, and 1 study (Gupta et al. 2016) was related to our research question 2. No studies were included in a quantitative synthesis.

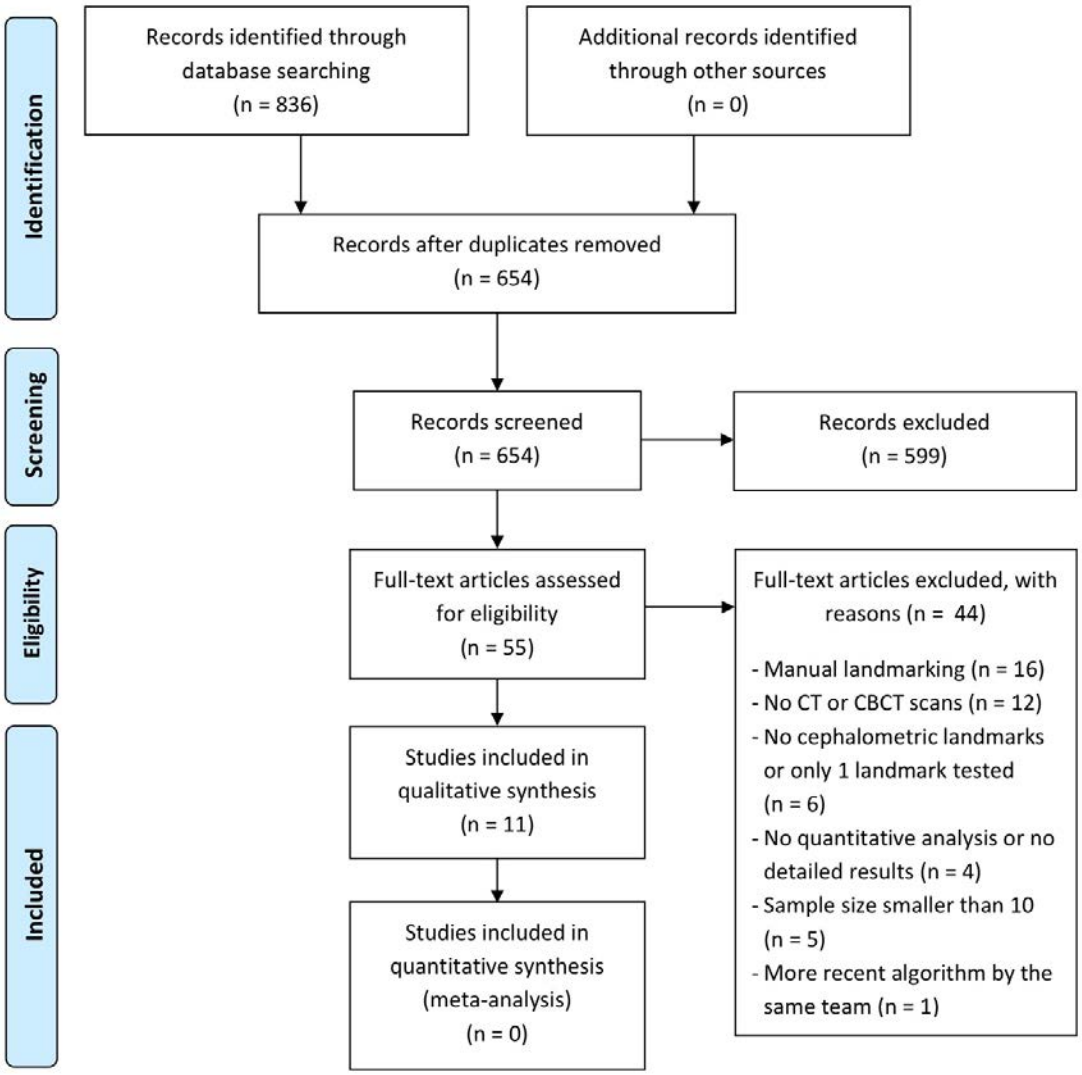


Figure 17: Flow chart of data searches using PRISMA guidelines.

2.4.2. Study characteristics

All of the selected studies for the qualitative analysis were published in English between 2014 and 2019 and were based on a retrospective selection of in vivo CBCT or CT scans. The sample size of the training dataset ranged from 24 to 201 images, and from 18 to 77 images for the test dataset.

None of the articles provided detailed descriptions of sample characteristics (e.g. gender, age, inclusion criteria, exclusion criteria, main craniofacial characteristics) nor provided calculation of the sample size. Four studies reported the use of a random method for population selection, but none reported

the details of the randomization. A various number of skeletal and dental landmarks were localized, ranging from 2 to 33 points of cephalometric interest.

Three main types of algorithms were used for the automatic 3D landmarking: knowledge-based, atlas-based and learning-based methods. A synthesis of the principles, advantages and limitations of these algorithms is provided in Table 3. The computational power needed and running time of the algorithms was stated in only 2 studies (Montúfar et al. 2018; O'Neil et al. 2019). For all studies, the reference standard used was manual landmarking. The reference landmarks were usually obtained by calculating the mean of landmarks provided by several observers. Summaries of the descriptive characteristics of the included articles are provided in Table 4 for research question 1 and Table 5 for research question 2.

Table 3: Principles, advantages and limitations of the algorithms used in the included articles

General Method	Specific Method	Principle	Advantages	Limitations
Knowledge-based		1. Mathematical entities are associated with the landmark locations (e.g. peak, lowest point…) 2. The landmarks are automatically localized on each contour of the test image based on the definitions	- Applies the concept of manual plotting based on pre-agreed definitions	- Detection of contours is the vulnerable step - Landmarks placed on curved structures are hard to localize - Robustness can be challenged with severely deformed cases
Atlas-based		1. A reference image atlas is created, with landmarks placed manually by experts 2. One image of this atlas is automatically registered (fitted) on a test image 3. The landmarks are transferred on the test image	- Simple method with low amount of a priori information needed - Can be customized easily	- Atlases have to be accurate and match biological variations (for sex, age, ethnicity…) - Highly dependent on registration technique which can be computationally expensive - Robustness can be challenged with severely deformed cases
Learning-based	Active shape model (ASM)	1. The landmarks are placed manually by experts on the training sample images 2. A statistical model (mean shape) is created by scaling, rotating and translating the training shapes so that they correspond as closely as possible 3. The model is iteratively deformed to fit the test image and automatically localize the landmarks	- Well-described and thoroughly studied method - Low sensitivity to artifacts and noise in the image	- 2-dimensional technique - Needs large training sample size to match biological variations - Needs accurate training data - Robustness can be challenged with severely deformed cases
	Elastic Bunch Graph Matching (EBGM)	1. The landmarks are placed manually by experts on the training sample images 2. A large set of 2D features is derived from the training data, using image filtering 3. The landmarks are automatically detected on the test image based on a maximum correlation search between the test image and a graph representation extracted from the training images	- Does not need a large training sample	- 2-dimensional technique - Needs accurate training data - Sensitive to artifacts and noise in the image
	Random forest	1. The landmarks are placed manually by experts on the training sample images 2. Visual features are chosen and a multitude of decision trees is constructed from the training data 3. All these decision trees are automatically combined to vote for the most probable position of the landmarks on the test image	- Well-described and thoroughly studied method - Low sensitivity to artifacts and noise in the image	- Needs a large training sample size with artifacts and anatomical variations - Robustness can be challenged with severely deformed cases
	Deep learning	1. The landmarks are placed manually by experts on the training sample images 2. A deep neural network is trained with the sample data 3. The landmarks are automatically detected on the test image	- Can accommodate strong anatomical variations - Low sensitivity to artifacts and noise in the image - Highly dynamic research field	- Needs a very large training sample size with artifacts and anatomical variations - Needs accurate training data - Training phase is computationally expensive - Downsampling of images might be needed, which can increase uncertainty in results - Neural network parameters have to be determined empirically

Table 4: Summary characteristics of included articles – Research question 1

Article	Population and selection method	Acquisition voxel size	Number of landmarks tested	Index test – Automatic 3D landmarking			Reference standard – Manual landmarking		Main Results
				Algorithm used	Training dataset	Test dataset	Observers Repetitions	Intraobserver Interobserver results	Total mean difference ± SD between index test and reference standard
Shahidi et al. 2014	Random retrospective selection from private practice images, without "significant fractures or severe skeletal anomalies" Age 10-43	*Unknown*	14	Atlas-based method	n/a	20 CBCT scans	3 observers 2 times	Intraobserver – ICC = 0.89 Interobserver - ICC + 95% CI = 0.87 [0.82-0.93]	3.40 mm
Gupta et al. 2015	Random retrospective selection from postgraduate orthodontic clinic database "irrespective of age, gender and ethnicity"	Isometric 0.25-0.40 mm	20	Knowledge-based method	*n/a*	30 CBCT scans	3 observers 1 time	Interobserver – ICC > 0.9	2.01 ± 1.23 mm
Zhang et al. 2016	Retrospective selection Non-syndromic dentofacial deformity Skeletal Class II and Class III patients	CBCT scans : isometric 0.4 mm CT scans: 0.488 × 0.488 × 1.25 mm^3	15	Random forest-based method	41 CBCT scans 30 CT scans	41 CBCT scans (same as training - 5-fold cross validation)	1 observer 1 time	*n/a*	1.44 mm
Codari et al. 2017	Retrospective selection from private practice database "Adult healthy Caucasian women" Age 37-74	*Unknown*	21	Atlas-based method	n/a	18 CBCT scans	"Team of expert users" 1 time	Interobserver – ICC = 0.98	2.39 ± 1.73 mm
Zhang et al. 2017	Retrospective selection from private practice database Non-syndromic dentofacial deformities Even distribution between skeletal classes	CBCT scans: isometric 0.3 or 0.4 mm CT scan: 0.488 × 0.488 × 1.25 mm^3	15	Deep learning-based method	77 CBCT scans 30 CT scans	77 CBCT scans (same as training - 5-fold cross validation)	2 observers (on different images) 1 time	*n/a*	1.10 ± 0.71 mm
de Jong et al. 2018	Retrospective selection from orthodontic clinic database Non-syndromic cohort Age 16-54	Slice thicknesses between 0.3 and 1 mm	33	Elastic Bunch Graph Matching-based (EBGM) method	39 CBCT scans	39 CBCT scans (same as training - leave-one-out test)	1 observer 1 time	*n/a*	Mean error <2 mm for 10 landmarks
Montúfar et al. 2018	Random selection from public repository (Virtual Skeleton Database from the Medical Image Repository of the Swiss Institute for Computer Assisted Surgery)	Isometric 0.4 mm	18	Active shape model (ASM) + Knowledge-based method on subvolumes	24 CBCT scans	24 CBCT scans (same as training - leave-one-out test)	2 observers 2 times	Intraobserver: "12 of 18 landmarks reproducible within a 1.0-mm standard deviation"	2.51 ± 1.6 mm
Neelapu et al. 2018	Retrospective selection from postgraduate orthodontic clinic database "irrespective of age, gender and ethnicity"	Isometric 0.25-0.40 mm	20	Knowledge-based method	*n/a*	30 CBCT scans	3 observers 1 time	Interobserver – ICC > 0.9	1.88 ± 1.10 mm
Torosdagli et al. 2019	Retrospective selection from hospital database, including "congenital deformities fading to extreme developmental variations in CMF bones" and artifacts	Isometric 0.29 or 0.377 mm (before resampling)	9[a]	Deep learning-based method	50 CBCT scans	50 CBCT scans (same as training - 5-fold cross validation)	3 observers 2 times for 2 observers 1 time for 1 observer	Interobserver – ICC = 0.92	Mean error ≤0.5 mm for 8 landmarks
O'Neil et al. 2019	Retrospective selection from hospital database, containing "pathology, inclusive of haemorrhage, tumours and age-related change"	"Range of resolutions and slice thicknesses"	2[b]	Deep learning-based method	170 CT scans for training 31 CT scans for validation	20 CT scans	2 observers 1 time	Interobserver – mean=2.20mm / median=1.48mm	Observer A: 2.45 ± 2.53 mm Observer B: 3.49 ± 2.88 mm

CT, computed tomography; CBCT, cone-beam computed tomography; ICC, intraclass correlation coefficient

[a] Only mandibular landmarks [b] Only 2 out of the 22 studied landmarks had cephalometric interest

Table 5: Summary characteristics of included articles – Research question 2

Article	Population and selection method	Acquisition voxel size	Number of measurements tested	Index test – Automatic 3D landmarking			Reference standard – Manual landmarking		Main Results Deviations of measurements
				Algorithm used	Training dataset	Test dataset	Observers Repetitions	Intraobserver Interobserver results	
Gupta et al. 2016	Random selection from postgraduate orthodontic clinic database "irrespective of age, gender and ethnicity"	Isometric 0.25-0.40 mm	28 linear 16 angular 7 ratios	Knowledge-based method	*n/a*	30 CBCT scans	3 observers 1 time	Interobserver – ICC > 0.9	- Linear measurements – highest error 2.63mm; mean standard deviation between 0.35 and 2.46 mm - Angular measurements – highest error 2.12°; mean standard deviation between 0.46 and 2.40° - Ratios – highest error 0.03; mean standard deviation between 0.01 and 0.03

CBCT, cone-beam computed tomography; ICC, intraclass correlation coefficient

2.4.3. Risk of bias and applicability

Using a tailored QUADAS-2 tool, three studies were assessed as being at overall low risk of bias (Gupta et al. 2015; Gupta et al. 2016; Torosdagli et al. 2019) and two were at low concern regarding applicability (Zhang et al. 2017; Torosdagli et al. 2019).

Regarding patient selection, 7 studies showed an unclear risk of bias (Shahidi et al. 2014; Zhang et al. 2016; Codari et al. 2017; de Jong et al. 2018; Montúfar et al. 2018; Neelapu et al. 2018; O'Neil et al. 2019) and 8 showed applicability concerns (Shahidi et al. 2014; Gupta et al. 2015; Gupta et al. 2016; Codari et al. 2017; de Jong et al. 2018; Montúfar et al. 2018; Neelapu et al. 2018; O'Neil et al. 2019). This was mainly due to a lack of testing on random or consecutive patients and a lack of description of the population sample. Furthermore, 5 studies were at unclear or high risk of bias regarding the implementation of the reference standard (Zhang et al. 2016; Zhang et al. 2017; de Jong et al. 2018; Montúfar et al. 2018; O'Neil et al. 2019). They lacked a proper reference standard, or failed to report inter and intra-operator reproducibility results. Risk of bias and applicability assessment is summarized in Figure 18.

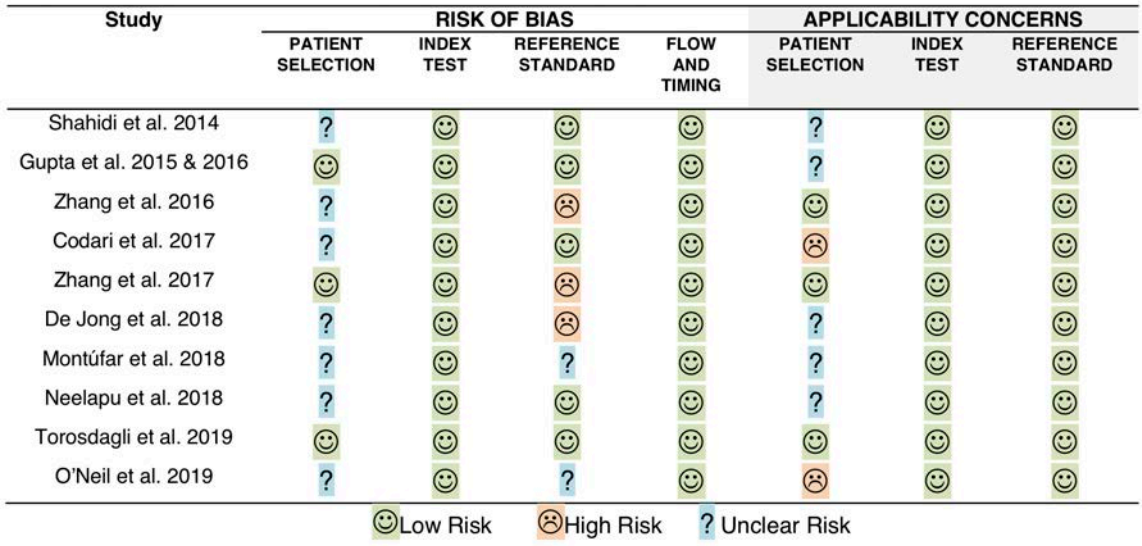


Figure 18: Bias and applicability assessment of included studies using tailored QUADAS-2 tool.

2.4.4. Results of individual studies: research question 1

For research question 1, the results were separated according to the method used for the automatization algorithm.

2.4.4.1. Knowledge-based methods

Knowledge-based methods use mathematical descriptions (e.g. peak, lowest point…) to localize the landmarks on the anatomical contours of the images. Two studies used this method. Detailed description of the algorithm and mathematical entities of the points were provided.

A first study by Gupta et al. (Gupta et al. 2015) tested 20 landmarks adapted from 2D cephalometry on a dataset of 30 CBCT scans. The initialization of the algorithm was based on the automatic search of a "seed point" using a template matching method on a segmented part of the images. A volume of interest (VOI) was defined around a point detected through distance vector from the "seed point". Then, landmarks were detected on the contours identified on the anatomical structures of VOI. The overall mean error was 2.01 mm (standard deviation 1.23 mm).

Neelapu et al. (Neelapu et al. 2018) aimed at improving the results and robustness of the aforementioned method. After segmentation of the images, algorithm initialization was based on the automatic localization of the mid-sagittal plane using symmetry features of the skull. The image was then cropped into four quadrants and landmarks were detected on the anatomical contours. The algorithm showed slightly better results than the previous study, and was said to be more robust for deformed cases. The overall mean error was 1.88 mm (standard deviation 1.10 mm) for the 20 landmarks.

2.4.4.2. Atlas-based methods

Atlas-based methods use an atlas of one or more reference images, with landmarks manually placed by experts. In order to localize landmarks on a new image, one of these reference images is automatically registered (fitted) on the test image and the landmarks are transferred. Two studies used this method.

Shahidi et al. (Shahidi et al. 2014) used 8 manually-annotated CBCT scans to generate the head atlas. The algorithm was tested with 14 landmarks on 20 CBCT scans. Depending on the age of the subject, one image of the atlas was automatically selected and fitted on the test image. The algorithm used feature and voxel similarity-based registration before scaling, rotation, and translation of the test image. The overall mean error was 3.40 mm.

Codari et al. (Codari et al. 2017) tested the automatic localization of 21 landmarks on 18 CBCT scans of healthy adult Caucasian women. One manually-annotated CBCT scan was used for the head atlas. After

automatic segmentation of the test image using k-means clustering, the atlas image was automatically fitted on the test image, using first an affine (linear) intensity-based image registration technique and then an elastic (nonlinear) one. The overall mean error was 2.39 mm (standard deviation 1.73 mm).

2.4.4.3. Learning-based methods

Learning-based methods include various methods which rely on a training sample of images. Two sub-types can be described: statistical and machine learning methods. Statistical methods (active shape model and Elastic Bunch Graph Matching) correlate a shape with deformation modes, or a graph representation, extracted from the training images, with the test image. Machine learning methods (random forest and deep learning) use the training data in order to learn where to localize the landmarks without being explicitly programmed to perform this task.

Montúfar et al. (Montúfar et al. 2018) used a combination of learning-based and knowledge-based methods. First, 2 active shape models (ASM) were trained on digitally reconstructed 2D radiographs for a holistic automatic 2D landmark approximation. Then, the 3D coordinates of the points were computerized and segmentation of the images' sub-volumes was performed around the points. Finally, a knowledge-based method was used to localize the landmarks precisely on the anatomical contours. The ASM was trained on 24 CBCT scans, and the overall localization results were tested on the same set of images (leave-one-out test). The mean error was 2.51 mm (standard deviation 1.6 mm). In terms of processing time, Montúfar et al. method was compared to Gupta et al. and Codari et al.. Reported processing times were 49,126.25 and 2,892.2 seconds, respectively.

De Jong et al. (de Jong et al. 2018) used an Elastic Bunch Graph Matching-based (EBGM) method. The training dataset consisted of 39 CBCT scans which were manually segmented and landmarked once by one operator. A total of 33 landmarks were localized. The segmented skulls were projected on a 2D plane and a large set of features was derived from this data. EBGM method was used to search for a maximum correlation between the test image and a graph representation extracted from the training images. A leave-one-out test was used to evaluate the algorithm, and 10 landmarks had a mean error inferior to 2 mm.

Zhang et al. (Zhang et al. 2016) used a random forest method to automatically localize 15 landmarks on 41 CBCT scans in a 5-fold cross validation test. The method was based on a regression voting strategy, using image segmentation to remove uninformative voxels. Then, a partially-joint model was used to localize landmarks separately based on the coherence of their positions. The training dataset

consisted of 41 CBCT and 30 CT scans, which were labelled once by one experienced operator. The overall mean error of the automatic localization was 1.44 mm, with all the landmarks having a mean error inferior to 2 mm.

Three studies used a deep learning method. O'Neil et al. (O'Neil et al. 2019) used a shallow fully convolutional neural network (FCN) and atlas location autocontext in order to localize 22 landmarks in the head. Atlas location autocontext was described in this work as "iteratively feeding the coordinate in atlas space, according to the output of a model, to a subsequent model". Two of these 22 landmarks had a cephalometric interest. A total of 170 CT scans were used for training, 31 for validation and 20 for testing. These images contained "pathology, inclusive of haemorrhage, tumours and age-related change". The data was manually labelled once by two observers, but the mean localization of the manual points was not used. The overall mean error of the automatic localization for the two points was 2.45 mm (standard deviation 2.53 mm) for observer A and 3.49 mm (standard deviation 2.88 mm) for observer B.

Zhang et al. (Zhang et al. 2017), in a second study, automatically localized 15 landmarks on 77 CBCT scans in a 5-fold cross validation test. Two fully convolutional neural networks (FCN-1 and FCN-2) with a U-Net architecture were used. FCN-1 was used to learn the displacement maps for multiple landmarks in order to model the spatial context information in the whole image. Then, FCN-2 performed bone segmentation and landmark localization using both FCN-1 results and the original image as input. The training dataset consisted of 77 CBCT and 30 CT scans. The overall mean error of the automatic localization was 1.10 mm (standard deviation 0.71 mm).

Torosdagli et al. (Torosdagli et al. 2019) used an adapted fully convolutional DenseNET network (also called Tiramisu network) for image segmentation, followed by an improved Zhang et al. (Zhang et al. 2017) U-Net network to localize sparsely-spaced landmarks. Then, a long short-term memory (LSTM) network was used to localize mid-sagittal closely-spaced landmarks near the "Menton" point. The training dataset consisted of 50 CBCT scans of mandibles including subjects with "congenital deformities fading to extreme developmental variations" and artefacts. The algorithm was tested on the same dataset using a 5-fold cross validation test. Eight out of the 9 mandibular landmarks tested were localized with a mean error inferior to 0.5 mm. The 9th one, "Pogonion", had a mean error of 1.55 mm.

We collected the detailed results of the mean errors and standard deviations for each cephalometric landmark. These mean errors were computed as mean distances (in mm) between the automatically-

detected test landmarks and the manually-detected reference landmarks (the latter being the mean of several observers, except for Zhang et al. (Zhang et al. 2016; Zhang et al. 2017), de Jong et al. (de Jong et al. 2018) and O'Neil et al. (O'Neil et al. 2019). Table 4 provides detailed results for the 19 most reported landmarks. The entire list of results can be found as Supplementary Table A3.

2.4.5. Results of individual studies: research question 2

For research question 2, the only study was performed by Gupta et al. (Gupta et al. 2016) following the same knowledge-based method as their other study. Linear, angular and ratio measurements were computerized using the manually-placed or automatically-placed landmarks. Then, the difference between the measurements was calculated as mean error. The unpaired t-test (95% level of significance) showed no statistically significant differences. For the linear measurements (Euclidian distance between two points), the highest error was 2.63 mm (mean standard deviation between 0.35 and 2.46 mm). For the angular measurements (angle between three points or two planes), the highest error was 2.12° (mean standard deviation between 0.46 and 2.40°). For the ratios (proportional measurements between two linear measurements), the highest error was 0.03 (mean standard deviation between 0.01 and 0.03).

2.5. Discussion

Our systematic review revealed that automatic landmarking of 3D craniofacial images is an active and current research field, as 5 out of 11 of our included studies were published in 2018 or 2019. Only one among the selected studies answered our research question 2 about the reliability of linear and angular 3D measurements obtained through automatic landmarking. This is quite surprising considering that diagnostic value of cephalometric analysis rests on linear and angular measurements, not merely on landmarks. Although these measurements are based on landmarks, overall measurement errors cannot be deduced systematically from landmark localization errors. Depending on landmark coordinate values, the overall measurement error can be reduced or increased, thus modifying the clinical significance of the results (Smektała et al. 2014; Gupta et al. 2016; Sam et al. 2018). Therefore, there is a lack of evidence about the diagnostic accuracy of automatic 3D cephalometry (Fryback and Thornbury 1991).

Concerning our research question 1, the best localization results were obtained by two studies that used a deep learning method to automatically localize the landmarks (Zhang et al. 2017; Torosdagli et al. 2019). More specifically, these two studies used fully convolutional neural network with a U-Net

architecture. Similarly, two of the best performing algorithms for automatic 2D cephalometry used a machine learning-based algorithm (Lindner et al. 2016; Vandaele et al. 2018).

These results need to be compared to those obtained through manual landmarking. Reproducibility studies of manual landmarking report variable results depending on the landmarks. Intra-operator results usually show mean differences smaller than 1 mm, and inter-operator variability ranges from less than 0.5 mm to more than 2 mm (Pittayapat et al. 2014; Smektała et al. 2014; Sam et al. 2018). At the moment, there is no clear threshold for clinical significance of inter-observer variability. Depending on the authors, the limit could be 0.5 mm, 1 mm, 2 mm or even more (Lagravère et al. 2010; Gupta et al. 2016; Sam et al. 2018). This questions the use of manual landmarking as the reference standard to test automated landmarking, but for now there is no other choice than to consider landmark localization by the mean of experts as the gold standard (Gupta et al. 2015; Lindner et al. 2016). A way to reduce uncertainty with this reference standard is to use the mean of manual landmarks obtained by several independent observers at different times.

When compared to the aforementioned body of literature, the localization results of the automated methods are very promising. Nonetheless, most of the algorithms were tested on a small set of cephalometric points or localized unconventional landmarks, as showed in Table 4 and Supplementary Table A3. This jeopardize the clinical application of most of these methods, which cannot be used to perform a complete 3D cephalometric analysis. In the detailed point-by-point results of the two best performing studies, some points show larger standard deviations than others. It is particularly noticeable in the results of Zhang et al. (Zhang et al. 2017) for points “Gonion Left” and “Gonion Right”, and in the results of Torosdagli et al. (Torosdagli et al. 2019) for points “Pogonion” and “Gnathion”. It is difficult to know what explains this phenomenon without detailed directional results for the errors. These landmarks are localized on curved structures with no clear boundaries, which are also known to be difficult to localize precisely in manual landmarking (Sam et al. 2018).

The performance of the learning-based algorithms entirely depends on the quality, size and variability of their training datasets (Lindner et al. 2016). The robustness of these algorithms needs to be more thoroughly tested in challenging and actual clinical sets, and time cost of the methods should be considered. These tests should primarily focus on the main target population of 3D cephalometry, difficult cases (e.g. patients with craniofacial syndromes, major asymmetries/craniofacial anomalies or undergoing orthognathic surgery) (Pittayapat et al. 2014; Kapila and Nervina 2015). As it has been done for automated 2D cephalometry, it would be interesting to gather a public and unbiased labelled set of images for the benchmarking of the algorithms (Wang et al. 2016). It would allow the training and

testing of the algorithms with a consistent evaluation method, thus helping the direct comparison between the results. In order to minimise the radiation dose of the patients, the algorithms should also be trained and tested on images acquired through low-dose protocols (SEDENTEXCT project 2012; Smektała et al. 2014).

Several studies showed risk of bias or applicability concerns regarding patient selection and implementation of the reference standard, mainly because the risks were assessed as unclear. Insufficient data has been reported in the included studies, therefore it cannot be ruled out that patients might have inappropriately been excluded or that the manual landmarking step might have failed to correctly detect the reference landmarks (Whiting 2011). Overall, some of the included studies might have yielded overoptimistic results. Interestingly, the study that provided the best results was also the only one that was assessed as being at overall low risk of bias and low concern regarding applicability (Torosdagli et al. 2019). However, it only focused on a set of mandibular landmarks and was validated on a rather small dataset.

The studies could have reported their results in other forms. Only three studies reported the percentage of points successfully located within a radius of 1 mm, 2 mm and 3 mm from the reference point. This data is needed to compute successful detection rates of the algorithms (Wang et al. 2016). Moreover, mean error might not be the most relevant result to assess distribution when error distributions are asymmetrical, as it is frequently the case with the algorithms used in the included studies. Median error and interquartile range should be used in that case (Codari et al. 2017). Finally, the error results were given as Euclidian distances in all included studies, without referring to the x- y- z-axis. Detailed directional results are necessary to identify the points that are prone to error in one plane more than the others and thus are of different clinical significance (Leonardi et al. 2008; Lagravère et al. 2010).

Finally, the reliability of landmarking does not necessarily translate into meaningful implications and clinically relevant results (Sam et al. 2018). The same limitation applies for now to manual 3D cephalometry (Pittayapat et al. 2014; Smektała et al. 2014). More studies on diagnostic thinking efficacy and therapeutic efficacy (Fryback and Thornbury 1991) of automatic 3D cephalometry are needed in order to know in which cases this technique is useful for diagnosis and treatment planning (SEDENTEXCT project 2012).

3 : L'Apprentissage profond

Les études basées sur l'utilisation d'algorithmes d'apprentissage profond pour la segmentation et le placement des points céphalométriques rapportent de meilleurs résultats que les précédentes méthodes, justifiant le choix de cette approche pour ce travail de thèse.

3.1. Définitions

Trois notions interdépendantes permettent de définir le champ de recherche de l'apprentissage profond (Figure 19). **L'intelligence artificielle**, décrite dès les années 1950, regroupe « l'ensemble des théories et des techniques mises en œuvre en vue de réaliser des machines capables de simuler l'intelligence humaine » (Dictionnaire Larousse 2022). C'est donc une notion très vaste, englobant un ensemble de concepts et de technologies. **L'apprentissage automatique**, décrit dans les années 1980, est un domaine de l'intelligence artificielle désignant les programmes informatiques capables d'effectuer une tâche ou de prendre une décision à partir de données, sans que leur comportement n'ait été explicitement programmé. Enfin, **l'apprentissage profond**, décrit dans sa forme actuelle dans les années 2000, est une sous-catégorie de l'apprentissage automatique reposant sur l'utilisation de réseaux de neurones à couches multiples.

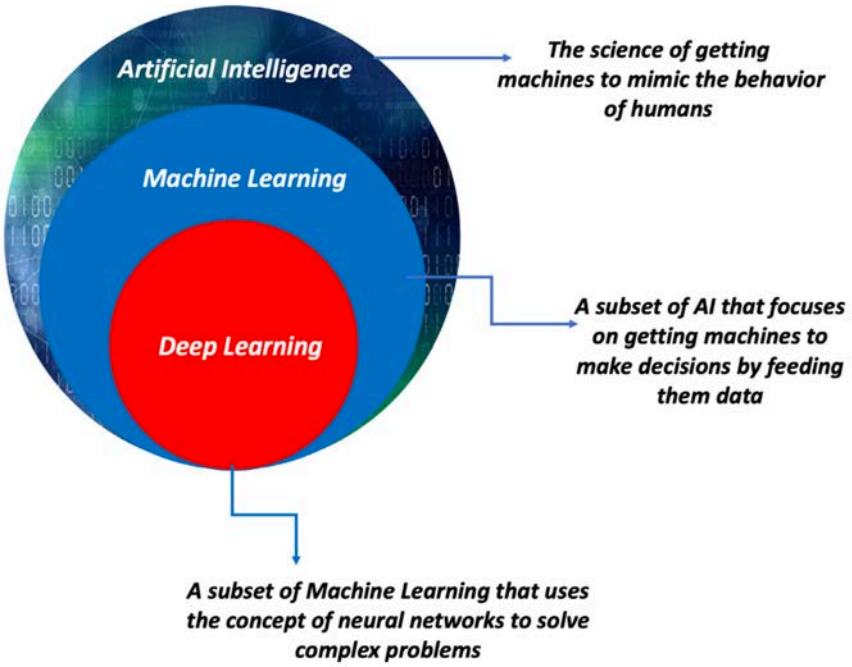


Figure 19 : Interdépendance des notions d'intelligence artificielle (« *artificial intelligence* »), apprentissage automatique (« *machine learning* ») et apprentissage profond (« *deep learning* »). Source: (Leonardi et al. 2021).

3.2. Grands principes

3.2.1. Apprentissage supervisé

Nous nous intéresserons dans ce travail plus particulièrement à l'apprentissage dit supervisé, qui constitue aujourd'hui l'utilisation la plus commune d'apprentissage automatique. Ce type d'apprentissage repose sur la création d'un modèle, entraîné sur un jeu de données, pouvant par la suite traiter de nouvelles données pour prédire un résultat. Les algorithmes sont décrits comme des modèles de classification si l'objectif est de prédire une valeur discrète (par exemple « spam » / « non spam », « chat » / « chien » / « éléphant », etc.), ou de régression si l'objectif est de prédire une valeur continue (par exemple un prix, un âge, etc.).

Imaginons que nous souhaitons construire un système pour classifier des images comme contenant un chat, un chien ou un éléphant (Figure 21). La première étape serait de collecter un grand nombre d'images de chats, chiens et éléphants qui seraient annotées par des opérateurs humains en fonction de leur catégorie (« chat », « chien » ou « éléphant ») et constitueraient notre base de données initiale. La démarche recommandée est ensuite de séparer cette base de données initiale en trois parties (Figure 20) :

1- la base de données d'entraînement, utilisée pour entraîner le modèle ;
2- la base de données de validation, utilisée pour le choix des différents paramètres du modèle (appelés hyper-paramètres, voir paragraphe 3.2.6) ;
3- la base de données de test, utilisée pour évaluer la performance du modèle entraîné.

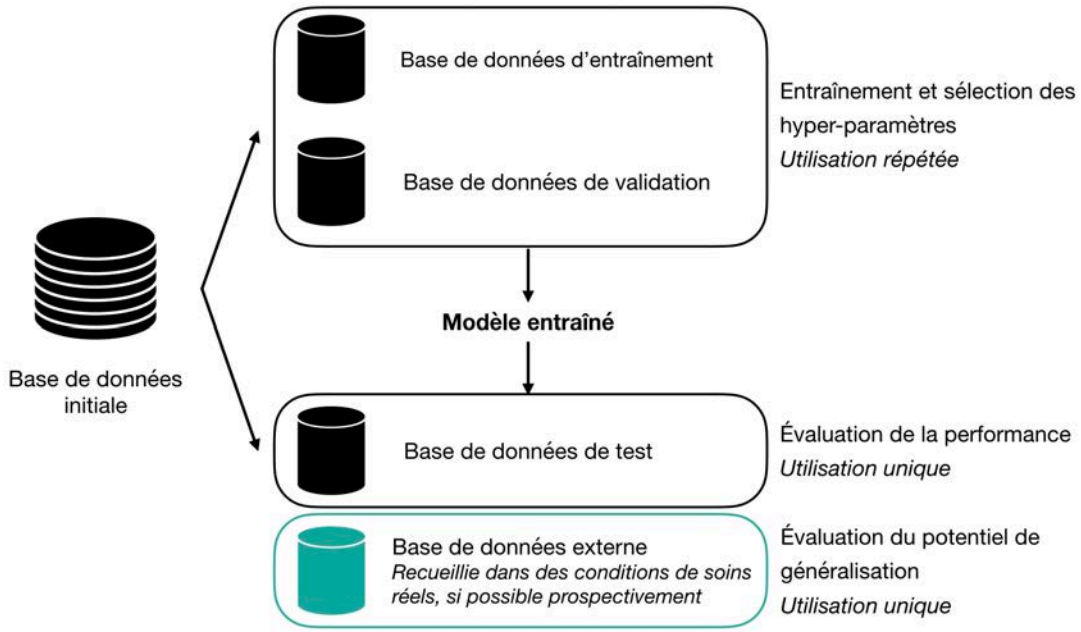


Figure 20 : Illustration des bonnes pratiques concernant l'utilisation des bases de données.

Pendant l'entraînement du modèle, les images de la base de données d'entraînement sont montrées à l'algorithme qui produit pour chacune un résultat sous la forme d'un vecteur de scores (un score pour chaque catégorie) (Figure 21A). L'objectif est de faire en sorte que la catégorie désirée (annotation « chat », « chien » ou « éléphant ») obtienne le score le plus élevé. Une fonction coût permet de mesurer l'erreur entre les scores désirés et les scores prédits (Figure 21B). Le modèle utilise ce résultat pour modifier ses paramètres internes ajustables dans le but de réduire la fonction coût et donc l'erreur de prédiction (Figure 21C). Ces différents paramètres ajustables, également appelés poids, peuvent être vus comme des boutons variateurs qui définissent le comportement du modèle. Après un premier ajustement des paramètres, les images d'entraînement sont à nouveau montrées au modèle, qui effectue de nouvelles prédictions dont l'erreur sera mesurée et utilisée pour ajuster à nouveau les poids du modèle. Ainsi de suite jusqu'à ce que la fonction coût se stabilise à un minimum, impliquant une stabilisation des poids du modèle. Le modèle est alors dit « entraîné », et ses poids sont fixés.

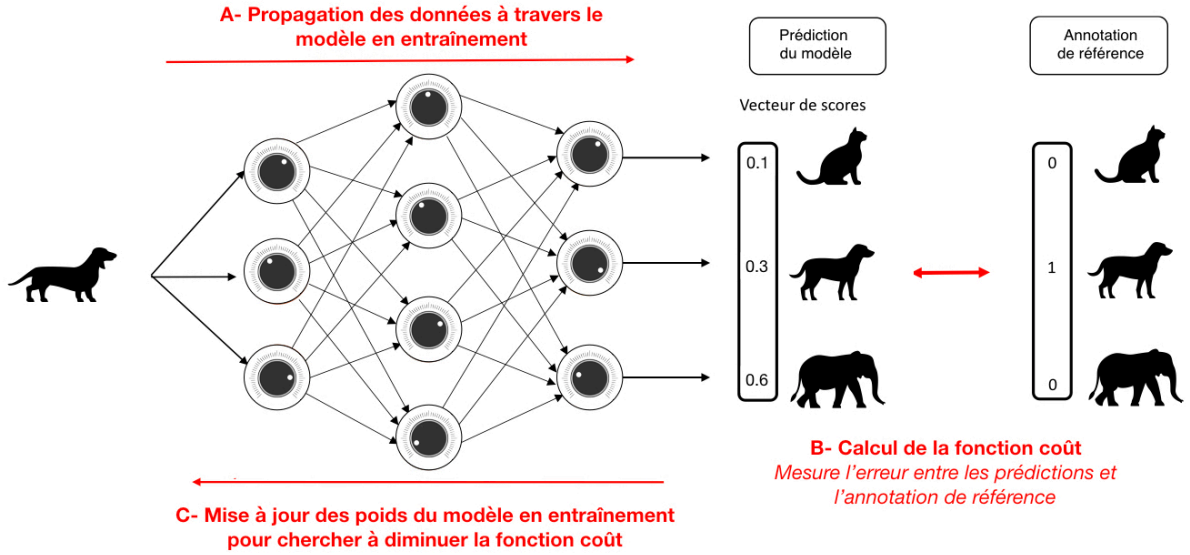


Figure 21 : Illustration schématique de l'entraînement d'un modèle d'apprentissage supervisé (ici un réseau de neurones) de classification.

Il est ensuite possible d'utiliser le modèle entraîné pour prédire la classification de nouvelles images (Figure 22). Une fois le modèle entraîné, celui-ci ne doit plus être modifié et l'évaluation de sa performance est effectuée en l'appliquant une seule fois à la base de données de test mise de côté : on montre les nouvelles images au modèle, qui va les analyser selon les poids précédemment fixés et en sortira des prédictions. Ces données ont été précédemment annotées par des opérateurs humains, mais sont complètement différentes de celles contenues dans les bases de données d'entraînement et de validation. Dans le domaine de la santé, plus le jeu de données de test comprend de données

hétérogènes et proches des conditions réelles d'utilisation (par exemple recueillies de façon prospective dans des conditions de soins réelles), plus la confiance dans l'évaluation du potentiel de généralisation du modèle sera grande et son déploiement clinique pourra être envisagé (Figure 20). A l'heure actuelle, la certification d'un algorithme d'apprentissage profond pour une utilisation clinique ne suit pas de voie réglementaire spécifique en Europe (marquage Conformité Européenne - CE) et aux États-Unis (approbation par la *US Food and Drug Administration* - FDA) (Muehlematter et al. 2021). Les algorithmes sont considérés comme des dispositifs médicaux et leur parcours réglementaire dépend du risque qui leur est attribué : plus ce risque est fort, plus des preuves importantes de sécurité, performance et fiabilité de l'algorithme sont à fournir. Pour les algorithmes considérés à haut risque, des études cliniques démontrant la sécurité et l'efficacité des dispositifs sont le plus souvent nécessaires (Muehlematter et al. 2021).

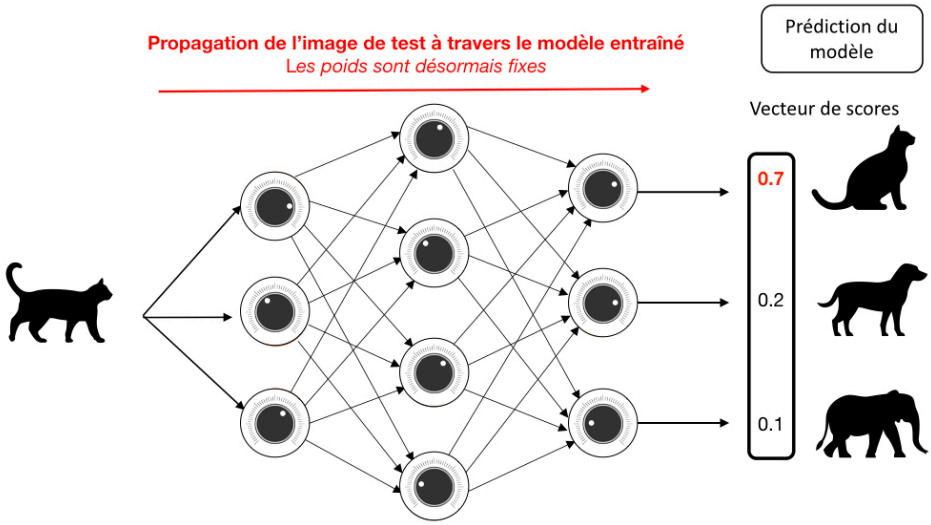


Figure 22 : Illustration schématique de la prédiction effectuée par un modèle d'apprentissage supervisé (ici un réseau de neurones) de classification préalablement entraîné.

3.2.2. De l'apprentissage automatique « conventionnel » à l'apprentissage profond

Dans le cadre des réseaux utilisés pour l'analyse d'image, chacun des poids du modèle est rattaché à une caractéristique que le modèle recherchera dans l'image analysée. Ces caractéristiques peuvent être plus ou moins abstraites en fonction de la complexité du modèle. Elles permettent de transformer l'entrée du modèle (par exemple une image) en une représentation appropriée pour que le programme informatique puisse y reconnaître des motifs.

L'apprentissage automatique dit « conventionnel » (ou apprentissage machine) repose sur la définition de ces différentes caractéristiques par des ingénieurs spécialisés, avant l'entraînement du modèle (Figure 23). Par exemple, deux caractéristiques couramment utilisées pour la détection d'objets sont

les caractéristiques HOG (« *histogram of oriented gradients* ») (Dalal and Triggs 2005) et Haar (Viola and Jones 2001). Des connaissances approfondies du domaine d'application et beaucoup de temps sont nécessaires pour construire et choisir ces caractéristiques, dont dépendra intégralement la qualité de prédiction du modèle (LeCun et al. 2015; Schwendicke et al. 2020). Chaque modèle a ainsi tendance à être optimisé pour une tâche précise.

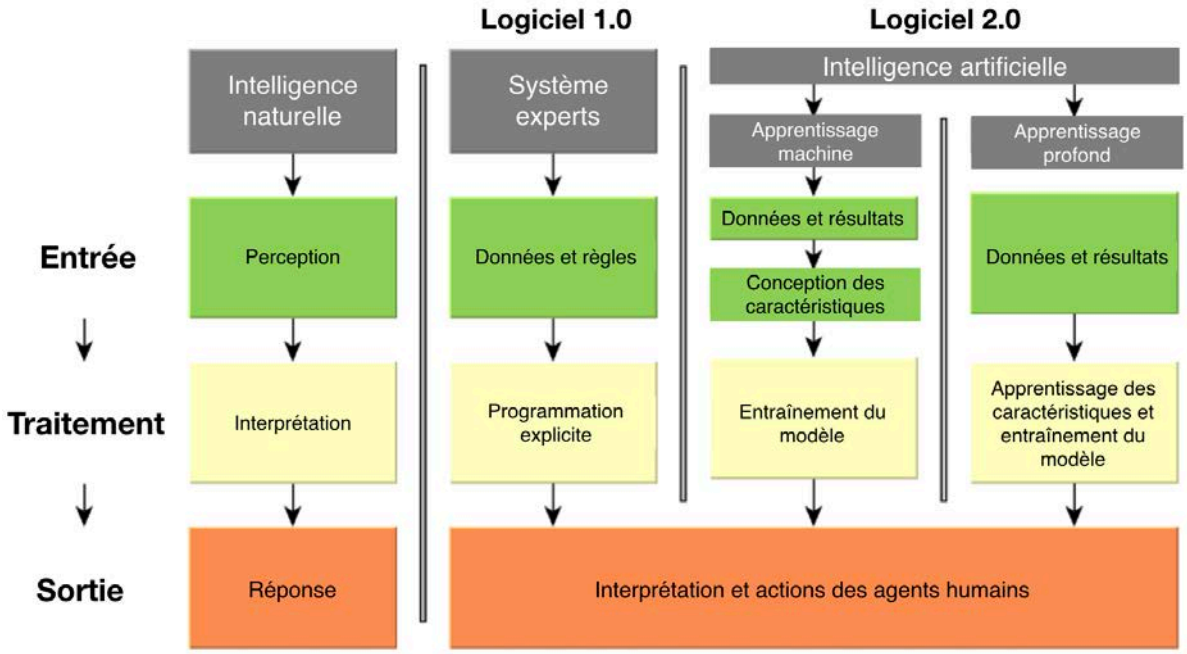


Figure 23 : L'intelligence naturelle est caractérisée par la perception, l'interprétation et la réponse. Les logiciels traditionnels 1.0 interprètent les données à partir de règles logiques explicitement programmées, tandis que les logiciels 2.0 utilisent les données et les résultats pour inférer les règles. Traduit d'après : (Schwendicke et al. 2020).

L'apprentissage profond a l'avantage majeur de proposer des modèles qui ont la capacité d'apprendre leurs caractéristiques dans le cadre de leur entraînement : les caractéristiques sont apprises et le modèle est entraîné dans une même étape, sans conception au préalable par des experts humains. (Figure 23). Chaque caractéristique est représentée par un « neurone », et les modèles comprenant de nombreuses couches de ces neurones s'appellent des réseaux neuronaux profonds (ou multicouches). Une architecture d'apprentissage profond est donc un empilement multicouche de « neurones » reliés les uns aux autres, qui font l'objet d'un apprentissage et calculent chacun des relations non linéaires (LeCun et al. 2015). Pendant l'entraînement du modèle, la mise à jour des poids du modèle (visant à minimiser la fonction coût) se fait par un algorithme de type descente de gradient stochastique, et le calcul de cette descente de gradient est effectué par un algorithme de rétro-propagation.

Cette approche permet une augmentation substantielle de la quantité de caractéristiques à analyser, un modèle d'apprentissage profond actuel pouvant présenter plusieurs centaines de millions de poids à ajuster et plusieurs milliards de connexions entre les neurones.

3.2.3. Réseaux neuronaux convolutifs

Les réseaux neuronaux convolutifs (« *convolutional neural networks »,* CNN) sont un sous-type de réseaux neuronaux, particulièrement adaptés à l'analyse d'images, dont l'organisation est inspirée par le fonctionnement du cortex visuel des animaux. Cette architecture a été rendue populaire en 2012 lors de la compétition ImageNet : l'application de CNN sur un jeu de données d'environ 1 million d'images tirées d'internet contenant 1000 annotations différentes a obtenu des résultats impressionnants, réduisant quasiment de moitié les taux d'erreurs des meilleures approches concurrentes (Krizhevsky et al. 2012). Ils représentent aujourd'hui l'approche dominante pour la grande majorité des tâches de reconnaissance et de détection automatiques (LeCun et al. 2015). L'entraînement de tels modèles a été rendu possible par d'importants progrès dans le matériel informatique - en particulier les cartes graphiques (« *Graphics Processing Unit* », GPU) – et dans le développement de librairies logicielles dédiées (par exemple *PyTorch* ou *Tensorflow*).

Le rôle des CNN est de réduire les images sous une forme plus facile à traiter pour le réseau, sans perdre des caractéristiques essentielles à l'établissement d'une bonne prédiction. L'architecture d'un CNN typique comprend deux types de couches qui se succèdent (Figure 24) :

- les couches de convolution (noyaux de convolution), qui traitent successivement de petites zones de l'image (initiale ou intermédiaire) et effectuent une opération de filtrage. Après l'opération de filtrage, une fonction d'activation est généralement appliquée sur le signal (par exemple une fonction ReLU, pour Unité Linéaire Rectifiée). Ces couches visent à détecter des caractéristiques locales de l'image, et produisent des images intermédiaires ;
- les couches de mise en commun (« *pooling* »), qui compressent l'information en réduisant la taille des images intermédiaires.

Les dernières couches du réseau sont souvent constituées de neurones intégralement connectés puis d'une couche permettant le calcul de la fonction coût.

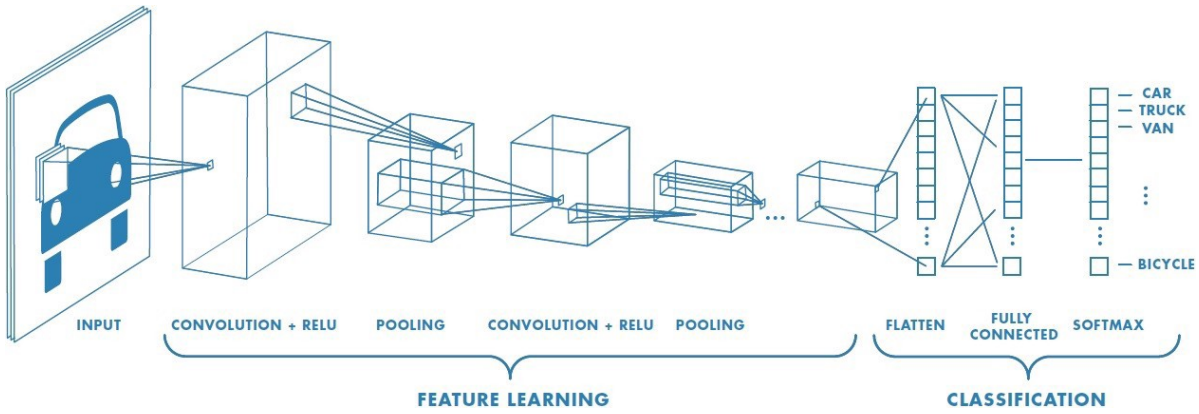


Figure 24 : Illustration de la succession des couches d'un réseau neuronal convolutif (CNN). Source : (Saha 2018 Dec 17).

Les premières couches des réseaux de type CNN apprennent généralement des caractéristiques extrêmement génériques, surtout lorsque les réseaux sont très profonds et entraînés sur de grands jeux de données (il est possible de visualiser ce que « voient » les différentes couches de plusieurs architectures CNN ici : https://microscope.openai.com/). Cette caractéristique permet de réutiliser des réseaux précédemment entraînés, en suivant des méthodes d'apprentissage « par transfert » : les poids des premières couches d'un réseau précédemment entraîné sont conservés, et uniquement les dernières couches du réseau sont « ré-entraînées ». Cette approche permet de réduire la quantité de nouvelles données nécessaires pour entraîner un réseau à effectuer une tâche précise.

3.2.4. Le réseau U-Net : l'approche de référence pour la segmentation d'images

Le réseau U-Net est une architecture de type CNN proposée pour la segmentation d'images biomédicales en 2015 (Ronneberger et al. 2015 May 18). L'objectif de cette architecture est d'obtenir des prédictions correctes tout en réduisant la quantité de données d'entraînement, puisqu'il est difficile d'obtenir des milliers de données annotées dans le domaine biomédical. Cette architecture intégralement convolutive repose sur deux parties successives : une partie contractante (encodeur) se rapprochant de la structure classique d'un CNN avec ses opérations de convolution et de mise en commun, et une partie expansive (décodeur) comprenant une séquence de convolutions et concaténations ascendantes (Figure 25). Dans le cadre de la segmentation d'images, l'image de sortie du réseau est un masque de segmentation, qui fait la même taille que l'image d'entrée et au sein duquel chaque pixel est classifié (par exemple comme « arrière-plan » ; « tissu 1 » ; « tissu 2 » ; etc.)

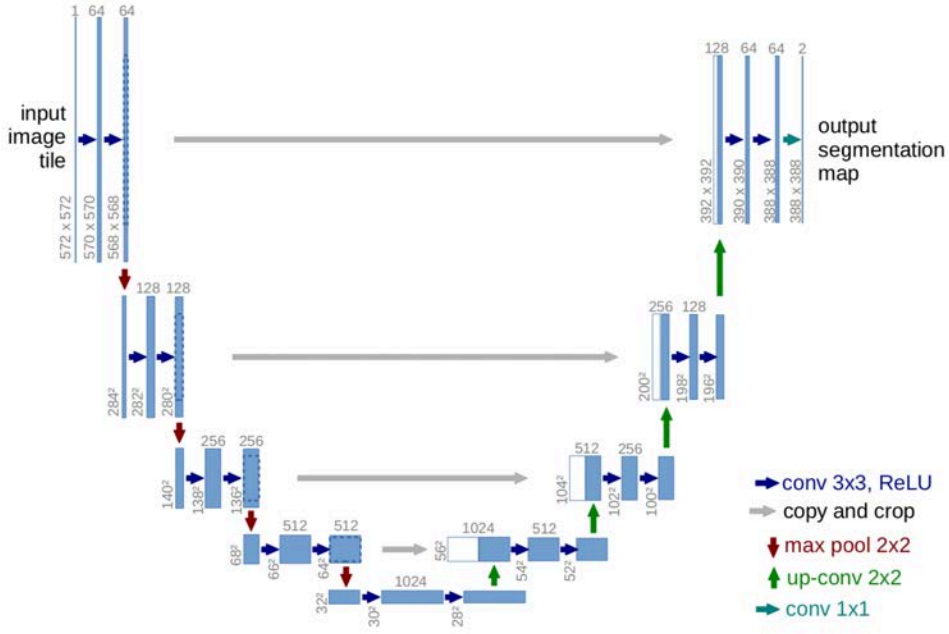


Figure 25 : Architecture du réseau U-Net originel. Source : (Ronneberger et al. 2015 May 18).

Rapidement, cette architecture a démontré son efficacité dans de très nombreuses applications, devenant la méthode de référence pour la segmentation d'images (Falk et al. 2019; Isensee et al. 2021). Initialement conçu pour des imageries 2D, le réseau a par la suite été adapté pour traiter les imageries 3D sous le nom U-Net 3D (Çiçek et al. 2016). Dans notre domaine, cette architecture U-Net 3D a été très largement utilisée pour la segmentation d'imageries dento-maxillo-faciales 3D (Tableau 1, page 18). Le volume de sortie est alors un volume de la même taille que celui d'entrée, avec chacun des voxels classifié en fonction du tissu auquel il appartient.

3.2.5. Les approches pour la localisation de points anatomiques

Si la segmentation des images repose sur des algorithmes de classification, les méthodes basées sur les CNN pour la localisation de points, et en particulier de points anatomiques, reposent le plus souvent sur des algorithmes de régression (Chen et al. 2021). Deux approches sont généralement proposées : (1) la régression directe des coordonnées des points ou (2) la régression de cartes de chaleurs.

Actuellement, les meilleures performances sont atteintes par les réseaux de régression de cartes de chaleurs, qui ont l'avantage de proposer une approche visuelle et bénéficient donc de l'efficience des réseaux CNN pour ce type de données (Zhang et al. 2020; Chen et al. 2021). Pour l'entraînement du réseau, des cartes de chaleur sont créées à partir de fonctions gaussiennes appliquées aux coordonnées des points : les pixels/voxels proches du point de référence ont des valeurs hautes, ces valeurs diminuant régulièrement en s'éloignant du point (Figure 26). Les résultats de sortie du réseau

(les prédictions) sont des cartes de chaleur du même type, qui représentent la probabilité, pour chaque pixel ou voxel de l'image, d'être tel ou tel point anatomique. Le traitement le plus simple est ensuite, pour chaque prédiction, de sélectionner le pixel/voxel ayant la probabilité la plus haute et d'en extraire ses coordonnées qui correspondront à la prédiction du point anatomique.

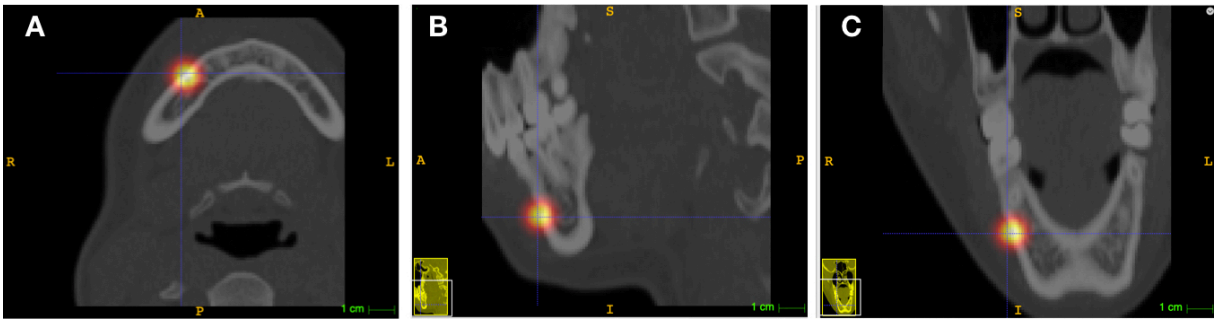


Figure 26 : Illustration d'une carte de chaleur utilisée pour l'entraînement d'un réseau de régression de point anatomique : A. Vue axiale ; B. Vue sagittale ; C. Vue frontale.

3.2.6. Mise en œuvre pratique : difficultés et réponses

Comme nous l'avons vu précédemment, les modèles d'apprentissage profond ont l'avantage de pouvoir apprendre leurs caractéristiques dans le cadre de leur entraînement. Ainsi, un réseau de type U-Net peut être vu comme une architecture générique qui fonctionnera pour de nombreuses tâches sans nécessiter de modifications majeures. Afin de l'utiliser pour une tâche précise, il « suffira » de l'entraîner sur une base de données correspondant à cette tâche. Cependant, la mise en œuvre pratique n'est pas si simple car de nombreux paramètres du réseau, appelés « hyper-paramètres », doivent être fixés manuellement par un opérateur humain.

La configuration de ces hyper-paramètres nécessite un haut niveau d'expertise et d'expérience, et de petites erreurs peuvent mener à des chutes drastiques de performance (Isensee et al. 2021). A l'heure actuelle, ces hyper-paramètres sont le plus souvent choisis de façon empirique en lançant des entraînements du modèle et en modifiant ses paramètres au fur et à mesure, à partir des résultats obtenus sur le jeu de données de validation (Figure 20). Isensee et al., en analysant les résultats de compétitions de segmentations biomédicales, ont montré que la même architecture (de type U-Net 3D) était à la fois utilisée dans les modèles avec les meilleurs et les moins bons résultats (Isensee et al. 2021). De plus, l'équipe a montré que les modifications apportées à l'architecture initiale du réseau U-Net n'offraient pas nécessairement de meilleurs résultats que l'architecture initiale correctement paramétrée. Cela démontre l'importance et la difficulté de la configuration manuelle des modèles.

Afin de répondre à ces difficultés, Isensee et al. ont proposé nnU-Net, une méthode configurant automatiquement les hyper-paramètres d'un réseau U-Net (2D ou 3D) pour la segmentation d'images biomédicales (Isensee et al. 2021). La configuration des hyper-paramètres se fait selon trois approches : certains hyper-paramètres sont fixés au préalable et sont toujours les mêmes, d'autres sont fixés à partir de l'analyse de la base de données d'entraînement (taille, intensité et modalité des images) et d'autres sont fixés à partir des résultats obtenus sur le jeu de données de validation. Sans aucune configuration manuelle, cette approche a surpassé la grande majorité de ses concurrents dans plusieurs compétitions et est aujourd'hui reconnue comme l'approche de référence pour la segmentation d'imageries biomédicales. Elle est disponible librement sur internet (https://github.com/MIC-DKFZ/nnUNet). Nous avons utilisé cet outil dans le cadre de notre étude sur la segmentation automatisée d'imageries CT-Scan dento-maxillo-faciales présentée en Chapitre 4.

Pour améliorer la répétabilité et faciliter l'adoption des résultats et des méthodes publiés, une approche répandue dans le champ de l'apprentissage profond est de partager librement sur internet l'architecture des réseaux utilisés (Mörch et al. 2021). Ainsi, ces réseaux peuvent être entraînés par d'autres auteurs sur leurs propres bases de données. Afin de faciliter le travail de configuration manuelle des hyper-paramètres, les méthodes ayant permis de fixer ces derniers doivent être les plus détaillées possibles par les auteurs dans leurs publications. Dans le cadre de notre travail, nous avons fait le choix d'utiliser exclusivement des réseaux disponibles librement.

3.3. Applications en odontologie

3.3.1. Domaines d'applications

Les premières applications de l'apprentissage profond et plus particulièrement des CNN dans le domaine dentaire ont été publiées en 2015. La quantité de publications a augmenté depuis cette date, sans connaître toutefois la même « explosion » que dans le reste des domaines médicaux (Figure 27). Il est probable que le domaine dentaire subisse un temps de latence de plusieurs années en comparaison avec d'autres domaines médicaux (dermatologie, ophtalmologie ou radiologie), qui pourrait s'expliquer par les difficultés à réunir de grandes bases de données et à préparer ces données pour obtenir des annotations de référence fiables (Schwendicke et al. 2019; Mörch et al. 2021).

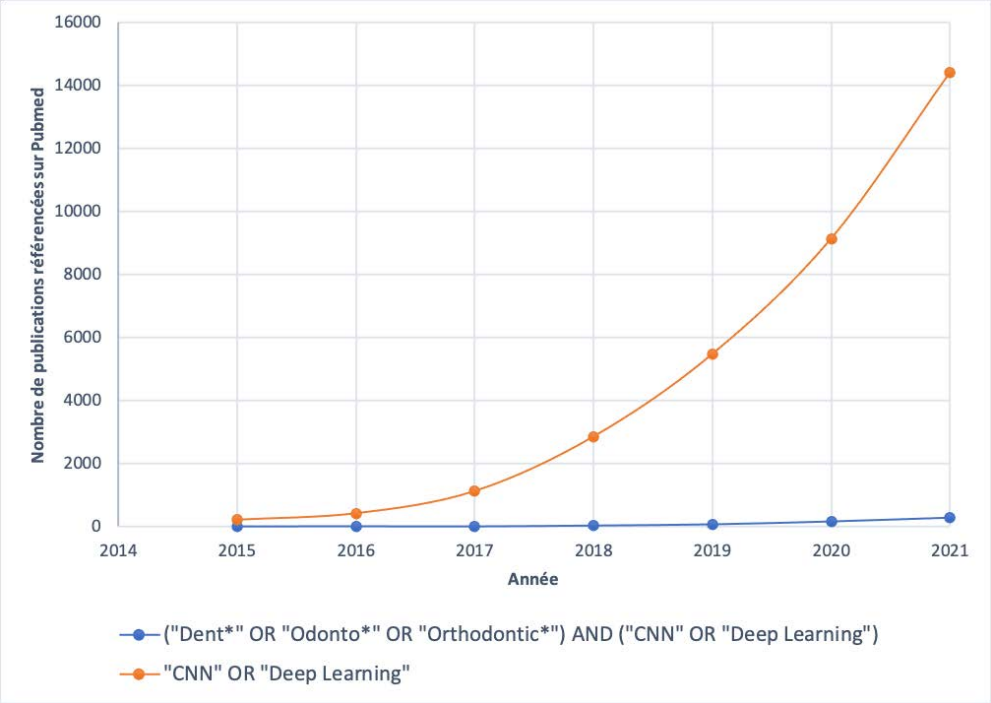


Figure 27 : Nombre de publications sur l'apprentissage profond ou les CNN référencées sur Pubmed dans le domaine dentaire et orthodontique (courbe bleue) et dans l'ensemble des domaines médicaux (courbe orange). Les requêtes utilisées sont indiquées en dessous des courbes.

D'après une récente revue de la littérature, les publications concernent de nombreux sujets dans les différents domaines de l'odontologie (Figure 28) (Mörch et al. 2021) :

- dentisterie générale : prévention sur la santé bucco-dentaire, détection du microbiote oral, prescription médicamenteuse, évaluation de l'indice de plaque, etc. ;
- cariologie : détection des lésions carieuses, classification du biofilm, reconnaissance des couches dentaires ;
- manifestations orales de maladies systémiques : détection de maladies rares ou de l'ostéoporose, prédiction du diabète ou de l'ostéonécrose, diagnostic du syndrome de Sjögren, etc. ;
- orthodontie : détection des points céphalométriques, classification des caractéristiques faciales, prédiction de l'intérêt d'une chirurgie orthognathique, évaluation de la nécessité d'un traitement orthodontique, etc. ;
- implantologie, parodontologie et chirurgie orale : détection de maladie parodontale ou de perte osseuse, reconnaissance et pronostic des implants dentaires, etc.
- endodontie : classification des traitements de racines et de la morphologie des racines, détection des lésions péri-apicales ;

- optimisation radiologique : amélioration de la qualité des images, segmentation et détection de points anatomiques, détection de lésions pathologiques ;
- médecine légale : estimation de l'âge, classification du sexe ;
- prothèse : optimisation des machines d'impression 3D et de fraisage, classification des préparations de couronnes, système d'aide à la décision ;
- cancers oraux : classification des images ou d'histologies de cancer oraux, détection protéomique et génétique des cancers oraux, etc.

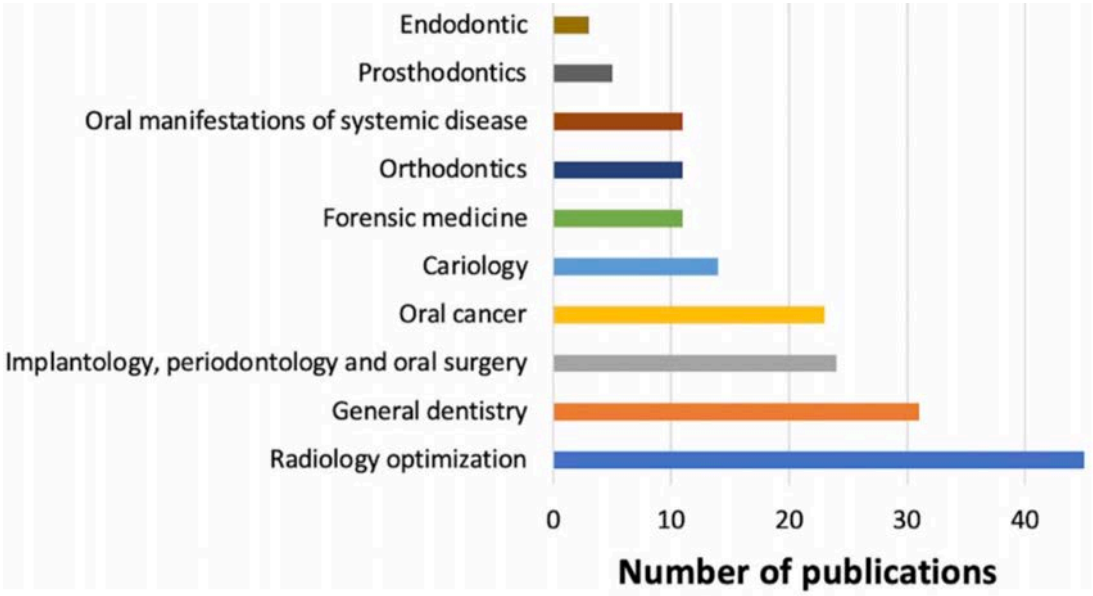


Figure 28 : Nombre de publications portant sur l'apprentissage profond classées en fonction des disciplines odontologiques. Source : (Mörch et al. 2021).

Dans le domaine de l'orthodontie en particulier, les trois principales utilisations des algorithmes d'apprentissage profond sont (1) la localisation des points céphalométriques sur des imageries 2D, (2) la localisation des points céphalométriques sur des imageries 3D et (3) l'aide au diagnostic et à la planification des traitements (Bichu et al. 2021).

3.3.2. Bases de données et annotations de référence

Dans le domaine de l'apprentissage profond, les données utilisées pour entraîner et évaluer les modèles sont essentielles. Les données et leurs méthodes de traitement décideront du potentiel de généralisation du modèle sur de nouvelles données jamais vues pendant l'entraînement du modèle. Les exemples de modèles montrant d'excellents résultats sur leur base de données d'entraînement mais n'arrivant pas à reproduire de tels résultats lorsqu'ils sont testés dans le « monde réel » sont nombreux. En dentisterie, les bases de données sont souvent petites et leur construction n'est pas

toujours détaillée (Tableau 1, page 18 et Tableau 2, page 24) (Schwendicke et al. 2020; Mörch et al. 2021; Schwendicke, Singh, et al. 2021) :

- les méthodes de sélection des données et la description de leurs caractéristiques ne permettent souvent pas de savoir précisément quelle est la population concernée par l'étude ;
- les résultats sont souvent évalués sur une partie de la base de données utilisée pour entraîner/évaluer le modèle (base de données de validation) et non sur un jeu de données mis de côté spécialement pour tester le modèle entraîné (base de données de test) ;
- les modalités d'annotation des références (nombre et caractéristiques des opérateurs, moyens utilisés, décisions en cas de désaccord entre plusieurs opérateurs…) sont souvent insuffisamment décrites.

Les bases de données et les méthodes d'annotations utilisées pour entraîner et tester les modèles d'apprentissage profond varient selon le domaine d'application. En fonction de l'objectif recherché, les experts se chargeant de l'annotation pourront par exemple segmenter les zones carieuses sur des acquisitions de trans-illumination ou des radiographies rétro-coronaires, ou alors simplement indiquer si une perte osseuse est présente ou non sur une radiographie rétro-alvéolaire (Figure 29). Les bases de données doivent être les plus hétérogènes possible afin d'augmenter le potentiel de généralisation du modèle. Par exemple, un modèle d'apprentissage profond entraîné avec des imageries issues de machines d'un seul fabriquant pourrait avoir des difficultés lorsqu'il rencontrera des imageries issues de machines d'un autre fabriquant. De même, un outil d'évaluation du risque carieux pour les enfants doit être testé sur des données d'enfants et non d'adultes. La source des données, les critères d'inclusion et d'exclusion des sujets et une description détaillée des caractéristiques démographiques et techniques de la base de données doivent être détaillés.

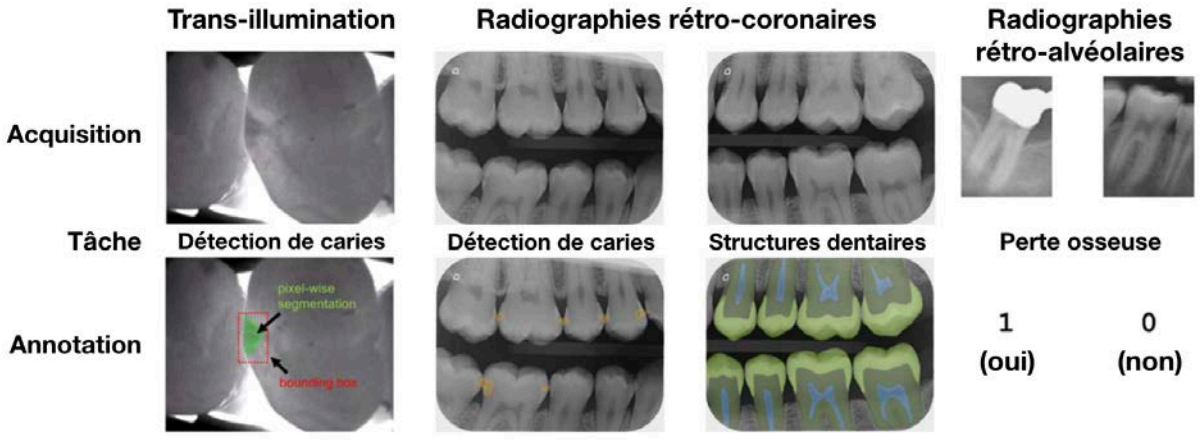


Figure 29 : Exemples d'acquisitions et d'annotations retrouvés en odontologie, en fonction de la tâche recherchée. Traduit d'après (Schwendicke et al. 2019).

L'annotation des données de référence est un enjeu important dans les applications dentaires de l'apprentissage profond, car il n'est souvent pas possible d'accéder à une réelle « référence » (par exemple une évaluation histologique *ex vivo* permettant de s'assurer de la présence ou de l'absence d'une carie). En conséquence, il est généralement nécessaire de demander à des experts humains d'annoter les données. Afin de diminuer l'aléa inter-opérateur, les opérateurs doivent être formés à la tâche et il est recommandé de demander à plusieurs opérateurs de labéliser les mêmes données. La façon dont l'annotation de référence est construite à partir de ces multiples annotations doit être détaillée : cela peut par exemple être un vote à la majorité (si 3 des 4 experts sont d'accord, l'annotation majoritaire est considérée comme « la vérité ») ou la moyenne de plusieurs annotations (typiquement dans le cas du placement de points céphalométriques).

3.3.3. Évaluation des résultats

L'évaluation des résultats des modèles s'effectue en comparant les prédictions avec les annotations de référence. Les méthodes utilisées pour évaluer ces résultats doivent être appropriées au sujet de recherche, et doivent reposer sur des critères de mesure pertinents.

Les disciplines techniques, qui mènent l'évolution de la recherche en apprentissage profond, ont l'habitude d'utiliser des critères d'évaluation qui sont souvent peu parlant pour les cliniciens. Par exemple, pour l'évaluation d'un modèle de segmentation le critère le plus communément utilisé est le DSC (voir paragraphe 1.3.1). Si ces critères permettent la comparaison de modèles dans le cadre de compétitions visant à classer différentes approches, ils ne sont pas toujours en cohérence avec des conséquences cliniques. En l'occurrence, le DSC n'est pas du tout adapté à l'évaluation de différences entre les formes. Ainsi, deux segmentation erronées avec le même DSC ne demanderont pas forcément le même travail pour être corrigées : une large erreur localisée sera plus rapide à corriger qu'une petite erreur retrouvée dans tout le volume (Nikolov et al. 2021).

Pour chercher à répondre à cette difficulté, un nouveau critère appelé surface DSC (sDSC) a été proposé (Nikolov et al. 2021). Celui-ci évalue le chevauchement entre les contours de deux segmentations, en se basant sur un seuil limite (par exemple 1 mm) pour faire la distinction entre les déviations cliniquement acceptables ou non (Figure 30). Le critère est calculé en faisant le rapport entre la somme des surfaces des portions acceptables et la surface totale de référence. En comparaison avec les critères classiques, le sDSC a été montré comme étant plus fortement corrélé au temps de retouche nécessaire pour corriger une segmentation en vue d'une utilisation clinique (Nikolov et al. 2021).

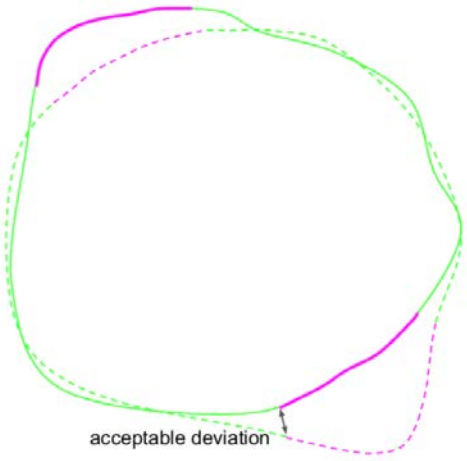


Figure 30 : Illustration du calcul du surface DSC (sDSC). Contour continu : référence ; contour en pointillé : prédiction ; flèche noire : seuil limite de déviation cliniquement acceptable. Vert : portions acceptables ; Rose : portions non acceptables. Source : (Nikolov et al. 2021).

Dans le cadre de l'évaluation du placement de points céphalométriques, les critères d'évaluation communément utilisés reposent sur le calcul de la distance euclidienne (« erreur ») entre les points de référence et les prédictions (erreur radiale moyenne et/ou taux de détections réussies) (Wang et al. 2016). Les auteurs ont été encouragés à utiliser des critères présentant davantage d'intérêt clinique, en se basant par exemple sur l'évaluation de mesures céphalométriques ou de conclusions diagnostiques issues du placement de ces points (Schwendicke, Chaurasia, et al. 2021).

Idéalement, les critères d'évaluation ne devraient pas uniquement se baser sur des mesures de précision mais également sur d'autres critères plus proches de l'utilisation clinique finale du modèle. Par exemple, une évaluation prospective de l'impact clinique de l'utilisation d'un modèle d'apprentissage profond peut être envisagée dans le cadre d'un essai clinique randomisé (Schwendicke et al. 2020).

3.3.4. Enjeux éthiques et recommandations de bonnes pratiques

Les questionnements éthiques relatifs à l'utilisation de modèles d'apprentissage profonds en odontologie sont nombreux. En particulier, trois difficultés éthiques principales ont été relevées dans une récente revue de la littérature (Mörch et al. 2021) :

- la prudence : l'anticipation et si possible l'évitement des conséquences négatives des outils, nécessitant par exemple que les opérateurs connaissent les limites des modèles utilisés ;
- la confidentialité : la constitution des bases de données dans le respect de la confidentialité ;

- la responsabilité : le développement de l'utilisation de modèles d'apprentissage profond ne doit pas réduire la responsabilité des opérateurs humains lorsqu'il font des choix médicaux. Il faut par exemple veiller à ce que le système ne devienne pas une référence pour des cliniciens peu expérimentés.

Les règles éthiques (accord éthique, consentement éclairé, etc.) et de protection des données (anonymisation, stockage, réutilisation, etc.) à respecter risquent de rendre le partage libre de données et la constitution de grandes bases de données multi-sites de plus en plus difficiles. Afin d'entraîner des modèles avec des données issues de plusieurs sites, une solution pourrait être le remplacement des entraînements centralisés par des entraînements fédérés sur plusieurs sites. Les modèles iraient alors où se trouvent les données, l'entraînement serait effectué localement puis les paramètres du modèles (les poids) pourraient être partagés avec l'ensemble des sites pour continuer l'entraînement sur un autre lieu (Schwendicke et al. 2019).

Afin d'aider à mieux appréhender les modalités de « raisonnement » des modèles d'apprentissage profond, le champ de « l'intelligence artificielle explicable » s'est développé ces dernières années. L'objectif est de permettre aux futurs utilisateurs de l'outil de savoir sur quels éléments s'est basé le modèle pour effectuer telle ou telle prédiction (Figure 31). Ces éléments d'explication permettent aux opérateurs de développer une analyse critique des résultats du modèle d'apprentissage profond, en s'éloignant du côté « boite noire » souvent utilisé pour décrire ces algorithmes. A l'avenir, il est probable que les modèles d'apprentissage profond doivent présenter ce type d'explications aux opérateurs pour l'obtention d'une approbation réglementaire à l'utilisation dans le domaine de la santé (Schwendicke et al. 2020).

Pour chercher à harmoniser et améliorer la qualité des études utilisant l'apprentissage profond dans la recherche dentaire, un groupe d'experts issus de l'association internationale pour la recherche dentaire (IADR) et du groupe de travail sur l'intelligence artificielle en santé commun à l'union internationale des télécommunication et à l'organisation mondiale de la santé (ITU/WHO Focus Group on Artificial Intelligence for Health (FG-AI4H) - https://www.itu.int/en/ITU-T/focusgroups/ai4h/) a récemment émis des recommandations de bonne pratiques (Schwendicke, Singh, et al. 2021). Les éléments présentés précédemment dans ce chapitre sont largement basés sur ces recommandations.

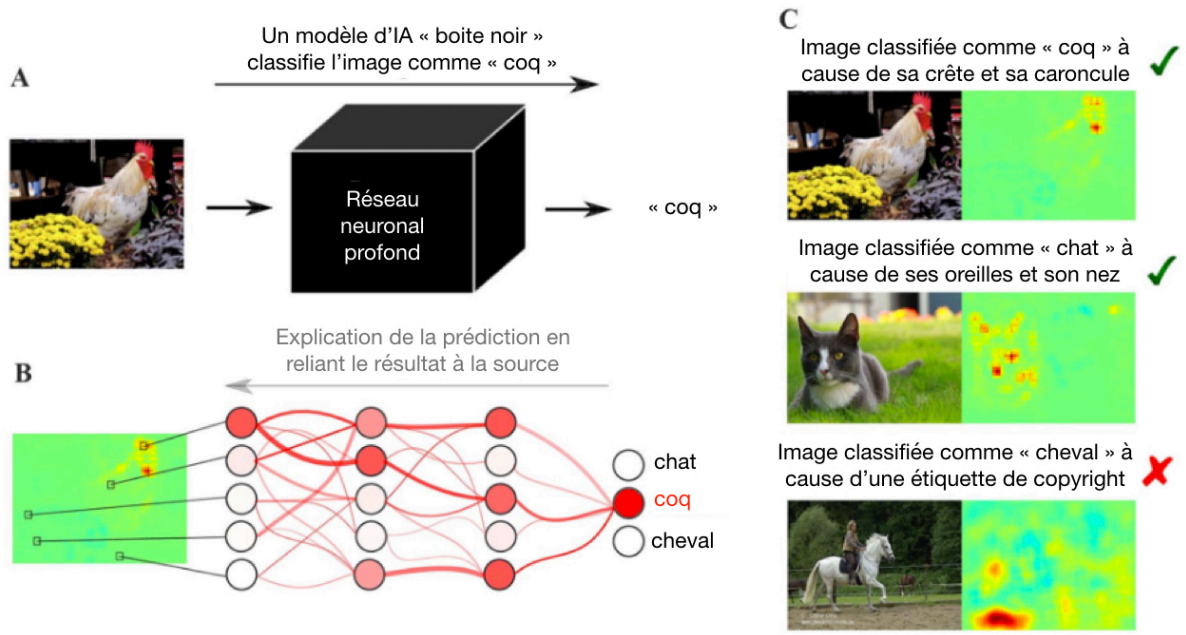


Figure 31 : « L'intelligence artificielle explicable » *:* A. Les modèles d'IA actuels sont considérés comme des « boites noires », faisant des prédictions sans expliquer comment ils y arrivent ; B. Les modèles « explicables » présentent sur une carte de chaleur quelles parties de l'image ont été importantes pour la décision de l'algorithme ; C. Cela permet de différencier les modèles prenant des décisions sur des éléments fiables (par exemple ici des éléments anatomiques) ou sur des éléments erronés (une étiquette de copyright). Traduit d'après: (Schwendicke et al. 2020).

Conclusion

L'automatisation du traitement des imageries dento-maxillo-faciales a connu ces dernières années des bouleversements majeurs avec le développement des modèles d'apprentissage profond. Les travaux publiés laissent entrevoir des applications dans l'ensemble des disciplines odontologiques, pouvant potentiellement offrir aux cliniciens une multitude de nouveaux outils. Ces outils automatisés seraient particulièrement bienvenus dans le domaine de la céphalométrie 3D, qui n'est pas utilisée en routine clinique en raison des difficultés rencontrées par les opérateurs pour effectuer la segmentation et le placement des points céphalométriques sur les imageries 3D dento-maxillo-faciales. Plusieurs travaux basés sur des modèles d'apprentissage profond présentent des résultats très prometteurs pour l'automatisation de ces deux tâches.

Cependant, un regard critique doit être porté sur l'applicabilité clinique des résultats publiés car la plupart des études souffrent de biais importants au regard des recommandations de bonnes pratiques. En particulier, les modèles sont testés sur un faible nombre d'imageries (moins de 50 pour la segmentation, mois de 10 pour le placement de points céphalométriques) dont les caractéristiques ne sont pas suffisamment décrites. Les méthodes d'annotation des données de référence ne sont généralement pas détaillées et les méthodes d'évaluation des modèles reposent sur des critères dont les implications cliniques ne sont pas toujours claires.

Au vu de ce contexte, notre travail a porté sur le développement et la validation de modèles d'apprentissage profond avec une attention particulière pour le respect des recommandations de bonnes pratiques. Notre base de données rétrospective était constituée de 453 imageries CT-Scan du massif dento-maxillo-facial, effectuées consécutivement dans un service de chirurgie maxillo-faciale. L'annotation des données de référence a reposé sur des experts industriels pour la segmentation et sur des experts cliniciens (dont la reproductibilité a été évaluée) pour le placement des points céphalométriques. Enfin, nous avons cherché à utiliser des critères d'évaluation originaux offrant des perspectives cliniques.

La deuxième partie de ce manuscrit détaillera notre travail portant sur la segmentation automatisée, alors que la troisième partie s'intéressera au placement des points céphalométriques. Enfin, nous illustrerons dans une quatrième partie l'applicabilité clinique de nos méthodes avec trois cas cliniques de patients pré-chirurgicaux.

PARTIE B :

Segmentation automatisée

Résumé

Cette deuxième partie s'intéresse au développement et à l'évaluation d'un modèle d'apprentissage profond pour la segmentation automatisée des imageries dento-maxillo-faciales 3D. La segmentation de ces imageries est nécessaire cliniquement afin de visualiser les tissus d'intérêts, placer les points céphalométriques ou encore planifier une chirurgie orthognathique.

Pour mener à bien ce travail, nous avons constitué une base de données de 453 imageries CT-Scan provenant de patients consécutifs opérés d'une chirurgie orthognathique au sein d'un service hospitalier. Dans le cadre de la planification chirurgicale de ces patients, les segmentations de référence ont été effectuées par un partenaire industriel et ont été utilisées pour la fabrication de guides chirurgicaux et de plaques d'ostéosynthèse sur-mesure. Nous avons utilisé la méthode librement accessible nnU-Net pour entraîner un modèle d'apprentissage profond à partir de ces données.

Nous présentons les résultats de l'évaluation quantitative de ce modèle sur une base de données de test de 153 CT-Scans, qui constitue la plus grande base d'imageries de test rapportée dans la littérature de ce domaine. Nos résultats sont comparables ou supérieurs à ceux des travaux publiés précédemment, malgré la tâche plus difficile à laquelle nous faisions face : les modèles de la littérature sont évalués sur 30 imageries au maximum et cherchent à segmenter moins de tissus.

Nous avons demandé à des experts industriels de valider ou non l'utilisation directe de 45 de ces segmentations automatisées pour effectuer une planification chirurgicale. Plusieurs segmentations n'ont pas été validées à cause d'une précision insuffisante au niveau des apex des dents. Sans prendre en compte les dents, les segmentations ont été validées pour une utilisation directe de planification chirurgicale pour 76 % des imageries. Au vu de la précision nécessaire pour ce type d'application, ces résultats sont très encourageants. Cette étude est présentée dans le Chapitre 4 et a été publiée en 2022.

Afin d'évaluer le potentiel de généralisation de ces résultats sur une base de données constituée d'imageries externes à notre service hospitalier, nous présentons une étude complémentaire et originale dans le Chapitre 5. Le modèle a correctement segmenté automatiquement ces 25 imageries, présentant d'excellents résultats quantitatifs.

4 : Segmentation automatisée de scanners cranio-faciaux pour la planification chirurgicale

Ce chapitre a fait l'objet d'une publication dans le journal *European Radiology* en 2022, sous le titre :

Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnU-Net framework

Gauthier Dot[1,2], Thomas Schouman[1,3], Guillaume Dubois[1,4], Philippe Rouch[1], Laurent Gajny[1]

[1] Institut de Biomecanique Humaine Georges Charpak, Arts et Metiers Paristech, Paris, France ;

[2] Universite de Paris, AP-HP, Hopital Pitie-Salpetriere, Service d'Odontologie, Paris, France ;

[3] Sorbonne Universite, AP-HP, Hopital Pitie-Salpetriere, Service de Chirurgie Maxillo-Faciale, Paris, France;

[4] Materialise, Malakoff, France.

DOI : https://doi.org/10.1007/s00330-021-08455-y

4.1. Abstract

Objectives To evaluate the performance of the nnU-Net open-source deep learning framework for automatic multi-task segmentation of craniomaxillofacial (CMF) structures in CT scans obtained for computer-assisted orthognathic surgery.

Methods Four hundred and fifty-three consecutive patients having undergone high-resolution CT scans before orthognathic surgery were randomly distributed among a training/validation cohort (*n* = 300) and a testing cohort (*n* = 153). The ground truth segmentations were generated by 2 operators following an industry-certified procedure for use in computer-assisted surgical planning and personalized implant manufacturing. Model performance was assessed by comparing model predictions with ground truth segmentations. Examination of 45 CT scans by an industry expert provided additional evaluation. The model's generalizability was tested on a publicly available dataset of 10 CT scans with ground truth segmentations of the mandible.

Results In the test cohort, mean volumetric Dice Similarity Coefficient (vDSC) & surface Dice Similarity Coefficient at 1mm (sDSC) were 0.96 & 0.97 for the upper skull, 0.94 & 0.98 for the mandible, 0.95 & 0.99 for the upper teeth, 0.94 & 0.99 for the lower teeth and 0.82 & 0.98 for the mandibular canal.

Industry expert segmentation approval rates were 93% for the mandible, 89% for the mandibular canal, 82% for the upper skull, 69% for the upper teeth and 58% for the lower teeth.

Conclusion While additional efforts are required for the segmentation of dental apices, our results demonstrated the model's reliability in terms of fully automatic segmentation of preoperative orthognathic CT scans.

Key points:

- The nnU-Net deep learning framework can be trained out-of-the-box to provide robust fully automatic multi-task segmentation of CT scans performed for computer-assisted orthognathic surgery planning.
- The clinical viability of the trained nnU-Net model is shown on a challenging test dataset of 153 CT scans randomly selected from clinical practice, showing metallic artifacts and diverse anatomical deformities.
- Commonly used biomedical segmentation evaluation metrics (volumetric and surface Dice Similarity Coefficient) do not always match industry expert evaluation in the case of more demanding clinical applications.

4.2. Introduction

Orthognathic surgery addresses congenital and acquired conditions of the facial skeleton by repositioning the jaws into a functional relationship in subjects presenting dentofacial deformities. It has been reported that up to 5% of the USA and UK populations could require orthognathic surgery (Borzabadi-Farahani et al. 2016). Surgical correction of craniomaxillofacial (CMF) deformities requires defining a specific surgical treatment plan for every single patient (Xia et al. 2009). In recent years, several teams have shown the reliability of computerized methods to analyze dentofacial deformities or to elaborate surgical treatment plans (Xia et al. 2009; Alkhayer et al. 2020). Virtual planning is usually performed using a CT scan or cone-beam CT (CBCT) scan of the patient's head (Alkhayer et al. 2020). The first step in the planning pipeline involves the extraction of structures of interest from the CT data by semantic segmentation. The choice of (CB)CT scan resolution, field of view (FOV) and anatomical structures to be segmented depend on what the planning is intended for. For the purpose of surgical guides or personalized implant manufacturing, the orthognathic surgery planning is based on high-resolution CT scans (with voxel size around 0.5*0.5*0.5mm^3) with full-head FOV (around 250mm). The following anatomical structures need to be segmented: upper skull and mandible, including the mandibular canals, upper and lower teeth (crowns and roots). Proper segmentation and delineation of these structures is known to be challenging due to factors such as large interindividual

morphological variations, tight connections between the structures, lack of contrast in joints and teeth apices, frequent presence of artifacts (orthodontic materials, fixation implants, dental fillings or crowns) (Torosdagli et al. 2017; Murabito et al. 2021).

Semi-automatic algorithms may be used to make segmentation less time-consuming (Wallner, Schwaiger, et al. 2019). Some of these methods provide high segmentation accuracy, yet are not fully automatic and therefore still require time-consuming manipulations by trained operators. In recent years, segmentations of medical images using deep learning algorithms have outperformed previously used algorithms. Deep learning algorithms might indeed be able to perform fully automatic segmentations (Nikolov et al. 2021). Several authors have developed specific deep learning-based models for automatic segmentation of the upper skull, mandibular bone, teeth or mandibular canal (Cui et al. 2019; Minnema et al. 2019; Qiu et al. 2019; Torosdagli et al. 2019; Chung et al. 2020; Jaskari et al. 2020; Kwak et al. 2020; Lian et al. 2020; Zhang et al. 2020; Murabito et al. 2021; Qiu et al. 2021; Wang et al. 2021). Most of these approaches relied on a U-Net convolutional architecture (Ronneberger et al. 2015 May 18), and yielded promising results, with reported volumetric Dice Similarity Coefficient (vDSC) between 90% and 95%. However, some limitations restrict their clinical applicability. None of them specified whether the dataset used for training and testing the algorithm included routine clinical cases, nor how the scans were selected. When the evaluation of the model on a hold-out test dataset was provided, the number of test scans was always less than 30. Moreover, none of the algorithms in the studies were used to segment all the structures of interest for computer-assisted planning of orthognathic surgery and personalized implant manufacturing, and only one work segmented bone and teeth separately (Wang et al. 2021). This calls into question the reproducibility and generalizability of previously published results, which are crucial factors for clinical validity (Nikolov et al. 2021; Schwendicke, Singh, et al. 2021).

Recently, the nnU-Net framework was proposed as an out-of-the-box tool which automatically configures itself in order to perform deep learning-based biomedical image segmentation (Isensee et al. 2021). This tool is publicly available and was shown to surpass most existing models on 23 public datasets used in international biomedical segmentation competitions. nnU-Net could be helpful for direct clinical applications, as it is open source and does not necessarily require expert knowledge to obtain competitive results. To our knowledge, the performance of this tool has not yet been evaluated for the segmentation of craniofacial hard tissue in a clinical context.

In this work, our main objective was to evaluate the performance of the nnU-Net framework for automatic segmentation of CMF structures in routine CT scans performed for computer-assisted orthognathic surgery.

4.3. Materials and Methods

4.3.1 Patient selection

Data were selected from a retrospective cohort of all consecutive patients having undergone orthognathic surgery in a single maxillofacial surgery department between January 2017 and December 2019. Patients referred to this center presented a wide variety of dentofacial deformities, came from various socioeconomic backgrounds and were ethnically diverse. Patients were considered for inclusion whatever dental deformity they presented, with no minimum age. Exclusion criteria were refusal to participate in the research (all patients were contacted by mail) and lack of industry-certified CT scan segmentations (see next paragraph: "Ground truth segmentation process"). Of the 473 subjects who underwent orthognathic surgery within the study timeframe, 4 refused to participate and 16 lacked industry-certified CT scan segmentations. 453 subjects (453 CT scans) were eventually included in our dataset (Figure 32). Mean in-plane pixel size of the CT scans was 0.45*0.45mm^2 and their mean slice thickness was 0.34mm. Most CT scans (n = 417) were obtained using a GE Healthcare Discovery (GEHC) CT750HD scanner with a tube current of 50mA, an exposure time of 730s and a tube voltage of 100kV. Scans were randomly distributed among a train/validation set (n = 300) and test set (n = 153). CT scans characteristics and CT machines are detailed in Table 6. This study was ethically approved by an Institutional Review Board.

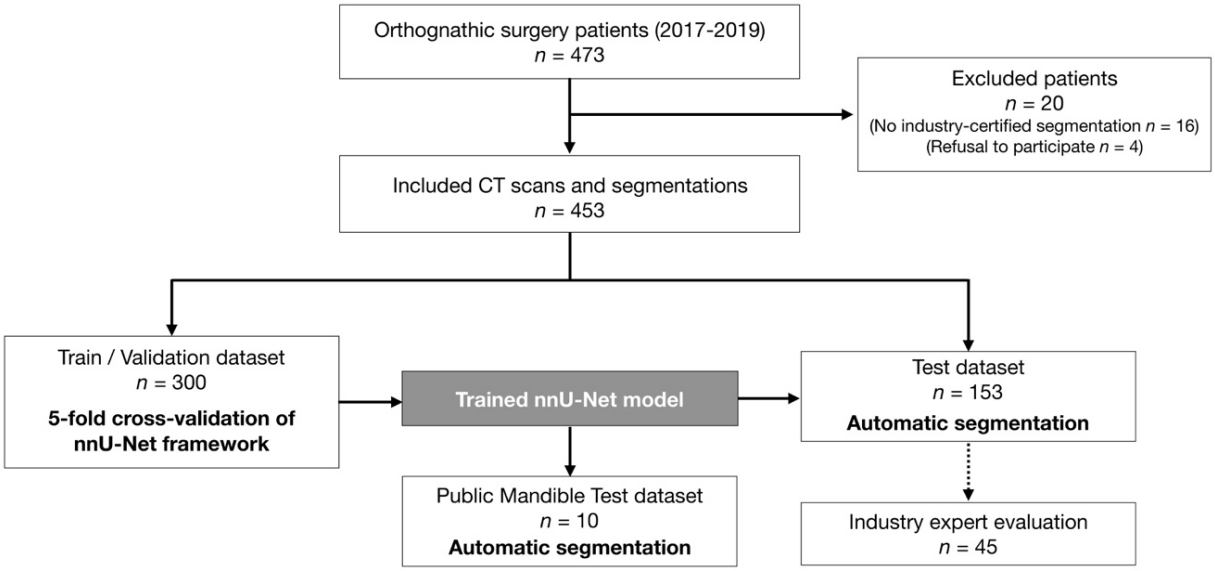


Figure 32: Data flow of the patient selection, training and evaluation process.

4.3.2. Ground truth segmentation process

All CT scans had been segmented prior to our study during patient treatment. Ground truth segmentations were used for diagnosis, computer-aided surgical planning and manufacturing of 3D-printed personalized surgical guides and fixation implants according to a certified internal procedure (Materialise). This industry procedure is confidential and cannot be fully described here. It implies semi-automatic segmentations, manually refined by a first operator [Step 1] before slice-by-slice verification for validation by a senior operator [Step 2] focusing mainly on the regions of interest: external surface of the bones, teeth and mandibular canals. Steps 1 & 2 are repeated until the segmentations are approved and certified for clinical use. This process results in 5 segmentation masks: (1) upper skull, (2) mandible, (3) upper teeth, (4) lower teeth and (5) both mandibular canals.

4.3.3. Public mandible test dataset

In order to assess the generalizability of our trained model, we tested it on a public dataset of 10 high-resolution CT scans from clinical practice (Wallner, Mischak, et al. 2019). These scans had the particularity of presenting complete mandibular bone structures entirely devoid of teeth. Ground truth manual segmentations of the mandible by two experts (A and B) were provided, which differed from our segmentations in that they were filled. Table 6 provides more details about this dataset. To the best of our knowledge, this is the only publicly available dataset of high-resolution Head CT scans with ground truth segmentation of one of the structures of interest for orthognathic surgery planning.

Table 6: CT scan characteristics and CT machines in the train/validation, test and public mandible test datasets.

	Train/Validation	Test	Public Mandible Test (Wallner, Mischak, et al. 2019)
Number of CT scans	300	153	10
Mean in-plane pixel size (mm^2)	0.44 * 0.44	0.45 * 0.45	0.45 * 0.45
Mean slice thickness (mm)	0.33	0.34	0.6
Number of scans by CT Machine			
GEHC Discovery CT750 HD	284	144	
GEHC Optima CT540 or CT660	7	5	
Siemens Sensation 64			10
Other CT Machine[1]	9	4	

GEHC: GE Healthcare. [1]GEHC Revolution CT, Philips Ingenuity Core, Siemens SOMATOM Definition AS, Toshiba Aquilion Prime SP

4.3.4. nnU-Net framework

The nnU-Net deep learning framework was used as an out-of-the-box tool, following instructions given by Isensee *et al.* (Isensee et al. 2021). Our raw train/validation dataset was used to automatically configure preprocessing, network architecture, training and post-processing pipelines. No modifications were made in setting the nnU-Net hyperparameters and data augmentation strategy. The training of 3D full resolution U-Net was performed on our train/validation set according to a 5-fold cross-validation strategy. After the end of the training pipeline, cross-validation result analysis showed that the model incorrectly labeled a few voxels as teeth in some scans displaying no upper and/or lower teeth. As a result, we implemented additional post-processing for teeth masks, consisting in removing all components smaller than an empirically-determined threshold. Finally, inference was performed on our test dataset as well as on the public mandible dataset (Figure 32). More details on the implementation of the deep learning framework are provided in Supplementary Materials B.

4.3.5. Evaluation metrics

Quantitative evaluation of the model performance was performed on our test set by comparing ground truth masks with predictions for each of the 5 segmentation masks. We followed best practice in evaluating model results, using both volume-based and surface-based metrics. Our main volume-based metric was the commonly used vDSC. Our main surface-based metric was surface DSC at 1mm (sDSC). Compared with classical metrics such as vDSC, sDSC, which was introduced recently, has been shown to be more strongly correlated with the amount of time needed to correct a segmentation for clinical use (Nikolov et al. 2021). We set our acceptable tolerance of sDSC at 1mm, as was done in recent international challenges in biomedical imaging. As in previous studies, an sDSC score of 95% was chosen as threshold value to consider variations between two segmentations as clinically non-significant. Additional quantitative metrics were computed after the common biomedical segmentation evaluations: Jaccard Coefficient, Volumetric Similarity, Average Surface Distance and Hausdorff distance.

4.3.6. Industry expert evaluation procedure

A random sample of our predicted masks for 45 subjects from our test dataset was sent to industry experts (Materialise) to be evaluated according to the 2-step validation process described above. The experts were blinded to our results and did not know that these segmentations were automatic, as they were evaluated among the flow of “classical” segmentations. Each segmentation mask was labeled “validated for clinical use” or “not validated for clinical use”.

4.3.7. Statistical analysis

Continuous variables are presented as mean ± standard deviation and categorical variables are expressed as numbers and percentages. vDSC and sDSC results are presented as percentages (%). The Wilcoxon test was used to compare vDSC and sDSC in scans obtained with GEHC CT750HD and scans obtained with other machines; p-values < 0.05 were considered significant. All data were analysed with Python (v.3.7) and RStudio software (v.1.3).

4.4. Results

4.4.1. Patient characteristics

In our database, patient mean age was 27 ± 11 years (minimum age 14, maximum age 66). The patients presented diverse anatomical deformities. 237 subjects (52.3%) exhibited skeletal class II (prognathic maxilla and/or retrognathic mandible), 163 subjects (36.0%) exhibited skeletal class III (retrognathic maxilla and/or prognathic mandible) and 165 subjects (36.4%) displayed asymmetry (clear asymmetry of the maxilla and/or the mandible evaluated on the 3D models). 89.6% of the CT scans (n = 406) showed metallic artefacts, in the form of orthodontic materials for 354 subjects (77.9%), metallic dental fillings or crowns for 193 subjects (42.6%) and fixation implants from previous orthognathic surgeries for 30 subjects (6.6%). Some subjects had no upper teeth (n = 4), no lower teeth (n = 1) or no teeth at all (n = 2). The public mandible test dataset included senior patients (63 ± 9 years) with no teeth at all. Table 7 summarizes patient characteristics in our datasets.

Table 7: Descriptive characteristics of the patients in the train/validation, test and public mandible test datasets.

Characteristic	Train/Validation (*n = 300)*	Test (*n = 153)*	Public Mandible Test (Wallner, Mischak, et al. 2019) (*n* = 10)
Age, mean ± SD, years	27 ± 11	26 ± 9	63 ± 9
Gender, no. (%)			
Female	178 (59.3)	83 (54.2)	5 (50.0)
Male	122 (40.7)	70 (45.8)	5 (50.0)
Skeletal deformity, no. (%)			
Class I	35 (11.7)	18 (11.8)	X[e]
Class II[a]	154 (51.3)	83 (54.2)	X[e]
Class III[b]	111 (37.0)	52 (34.0)	X[e]
Asymmetry[c]	112 (37.3)	53 (34.6)	1 (10.0)
Syndromic deformity	8 (2.7)	8 (5.2)	X[f]
Absence of teeth, no. (%)			
No upper teeth	4 (1.3)		
No lower teeth	1 (0.3)		
No teeth at all	1 (0.3)	1 (0.7)	10 (100)
Metal artifacts, no. (%)			
Orthodontic materials	232 (77.3)	122 (79.7)	
Metallic dental filling/crown	126 (42.0)	67 (43.8)	
Fixation implants[d]	17 (5.7)	13 (8.5)	
No metallic artifact	35 (11.7)	12 (7.8)	10 (100)

[a]Prognathic maxilla and/or retrognathic mandible. [b]Retrognathic maxilla and/or prognathic mandible. [c]Clear asymmetry of the maxilla and/or the mandible evaluated on the 3D models. [d]Fixation implants from a previous surgery. [e]Skeletal classification cannot be provided due to missing teeth and lack of information about the mandible's vertical position during CT scan acquisition. [f]No information provided with the database. SD, Standard Deviation.

4.4.2. Model performance

4.4.2.1. Quantitative evaluation on our test dataset

The mean results of vDSC and sDSC for each segmentation label of our test set are shown in Table 8, while Figure 33 shows the distribution of the results. The mean vDSC for all masks was 92.24 ± 6.19%. Without the mandibular canal results, the mean vDSC was 94.90 ± 0.91%. The mean sDSC for all masks was 98.03 ± 2.48%. Out of the 153 scans, 148 presented a mean sDSC for all masks which cleared the 95% limit for clinical significance. There were no statistically significant differences in both vDSC and sDSC when comparing scans obtained with GEHC CT750HD and scans obtained with other machines. Additional quantitative evaluation (Jaccard Coefficient, Volumetric Similarity, Average Surface Distance and Hausdorff distance) results for cross-validation and test datasets are provided in Supplementary Materials B.

Table 8: Mean vDSC and sDSC results on our test dataset (n = 153).

Metric, mean ± SD [%]	Upper Skull	Mandible	Upper teeth	Lower Teeth	Mandibular canal	Total
vDSC	96.22 ± 1.43	94.19 ± 1.62	94.83 ± 1.81	94.38 ± 2.32	81.59 ± 5.79	92.24 ± 6.19
sDSC	96.92 ± 3.08	97.92 ± 1.22	98.87 ± 1.18	98.53 ± 2.00	97.9 ± 3.51	98.03 ± 2.48

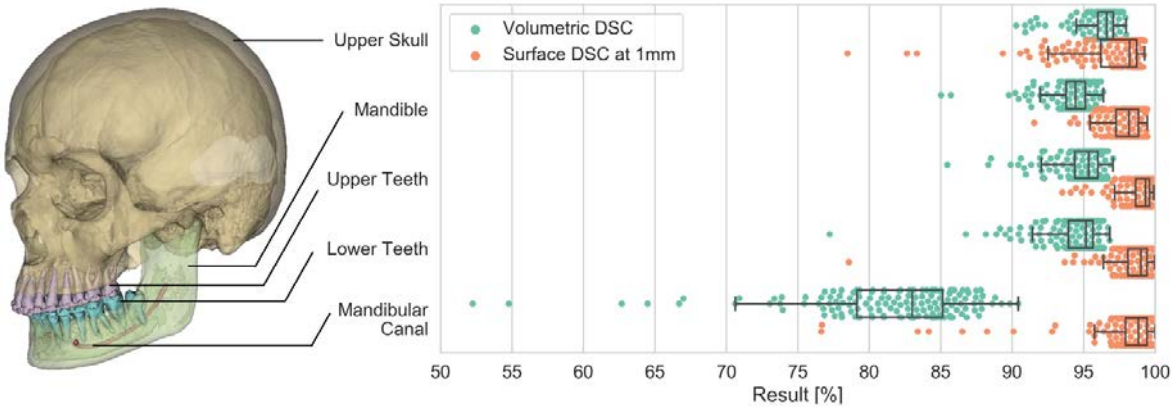


Figure 33: Left side: 3D model reconstructed from predicted segmentation masks (Upper Skull and Mandible with transparent overlay). Right side: distribution of vDSC and sDSC results in our test dataset (153 CT scans), for each segmentation mask (no result below 50%).

Four subjects representative of our test dataset and of the anatomic diversity of the database were chosen to illustrate our results (Figure 34, Figure 35). The most notable segmentation error made by the model on these subjects was the incorrect labeling of an upper deciduous tooth as a lower tooth in a patient suffering from craniofacial syndrome with several impacted teeth (Figure 34b, Figure 35b).

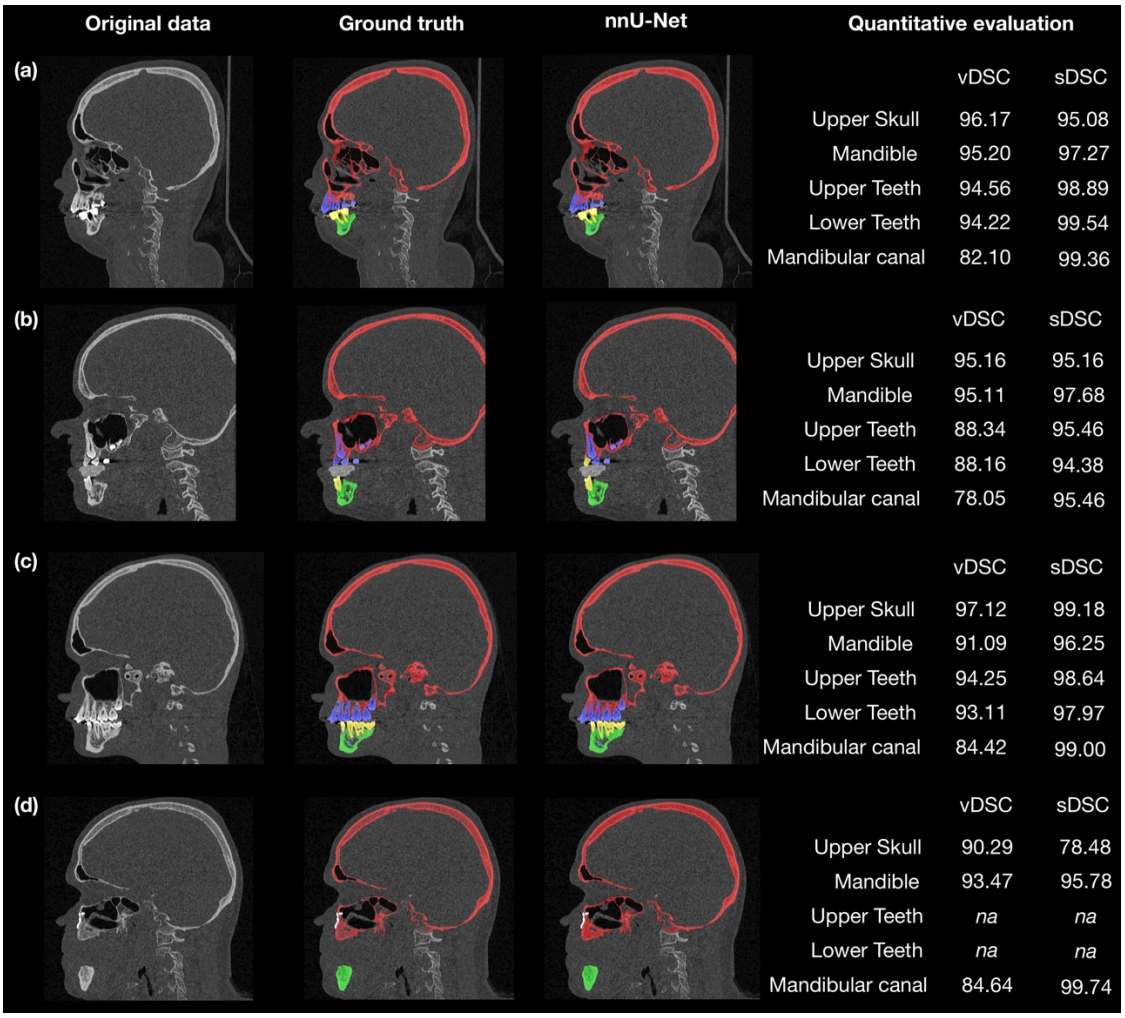


Figure 34: Four representative cases (a to d) showing sagittal slices of original data, ground truth segmentations, nnU-Net network-predicted segmentations and quantitative evaluation results. Red, upper skull; green, mandible; blue, upper teeth; yellow, lower teeth; cyan, mandibular canal.

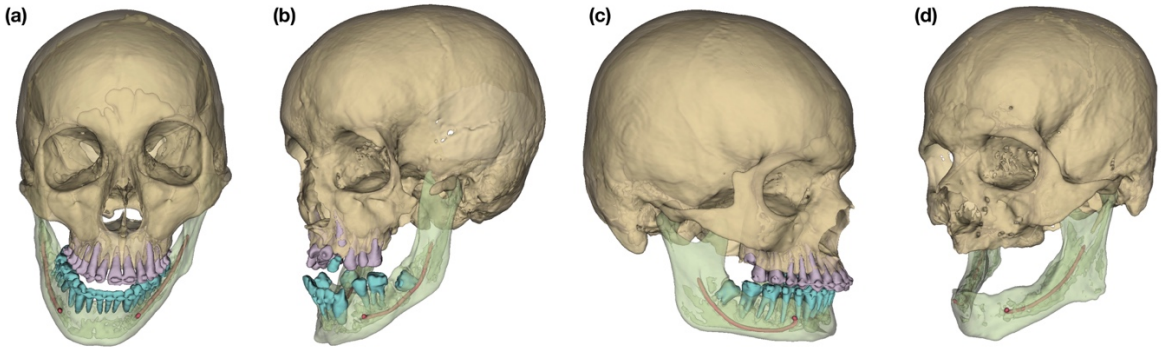


Figure 35: 3D surface models of segmentation results for 4 subjects representative of the anatomical diversity and of the challenges arising from our test dataset: (a) prognathic and asymmetric mandible; (b) craniofacial syndrome, with included and missing teeth; (c) retrognathic mandible; (d) no teeth and maxillary fixation implants from previous surgery (not segmented by the network).

4.4.2.2. Expert evaluation on our test dataset

The results of the industry expert validation process are shown in Table 9. Validation rates were about 90% for mandible and mandibular canals, 80% for upper skull and 60% for teeth. In total, 19 (42.2%) CT scans had all their segmentation masks validated. When excluding dental masks, 34 (75.6%) CT scans had their segmentation masks validated. Three CT scans out of the 45 (6.7%) did not show any metallic artefact, and had their 5 segmentation masks validated by industry experts. Comments appended to non-validated cases for bones mentioned small holes on the bony surface or under-segmentation of the anterior nasal spine. Reasons for rejection of mandibular canal masks were the inclusion of a few outlier voxels or the omission of an auxiliary canal during segmentation. As to teeth, reasons for non-validation were under- or over-segmentation of a few voxels of the apex (Figure 36).

Table 9: Results of industry expert evaluation on 45 random CT scans from our test set.

Result, no. (%)	Upper Skull	Mandible	Upper teeth	Lower Teeth	Mandibular canal	**Total**	**Total without teeth masks**
Validated	37 (82.2)	42 (93.3)	31 (68.9)	26 (57.8)	40 (88.9)	19 (42.2)	34 (74.6)
Not validated	8 (17.8)	3 (6.7)	14 (31.1)	19 (42.2)	5 (11.1)	26 (57.8)	11 (24.4)

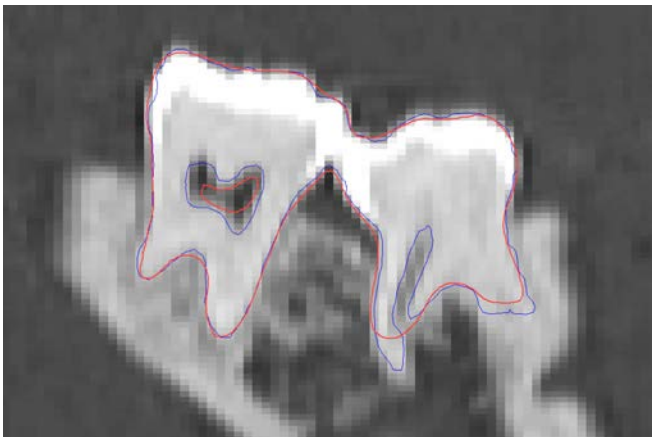


Figure 36: Representative lower teeth mask which was not validated by industry expert. Red line: predicted mask contour. Blue line: ground truth mask contour.

4.4.2.3. Quantitative evaluation on public mandible dataset

vDSC and sDSC results of the inference on the public mandible dataset are provided in Table 10. Most of the model's errors were located in the anterior part of the mandible, where the edentulous patients' alveolar bone was atrophic and extremely thin (Figure 37). One scan (mandible #10), performed on a subject with an endotracheal tube, produced low-quality results. Additional quantitative evaluation results for this dataset are presented in Supplementary Table B3.

Table 10: vDSC and sDSC results on public mandible dataset.

Metric, mean [%]	Operator ground truth	Mandible number									
		1	2	3	4	5	6	7	8	9	10
vDSC	A	91.72	85.12	89.29	89.38	92.16	90.74	92.15	88.71	92.66	61.19
	B	91.47	84.68	89.80	89.11	92.03	91.29	92.77	88.23	92.57	61.46
sDSC	A	97.10	92.96	95.62	90.56	98.03	94.12	97.29	90.84	99.10	63.86
	B	97.08	92.32	96.36	89.52	97.75	94.81	98.01	90.02	99.03	64.04

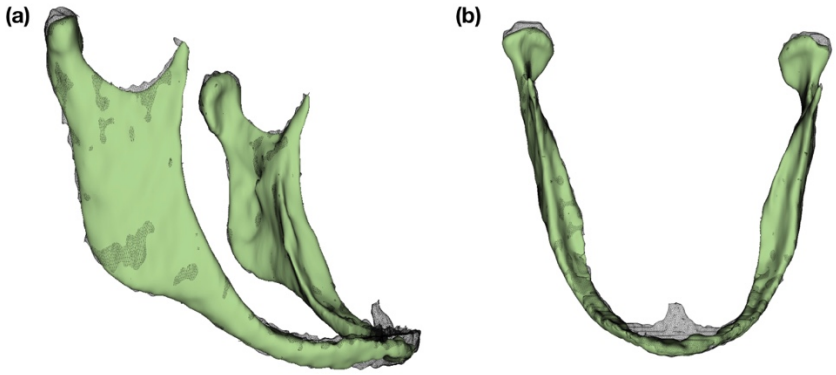


Figure 37: 3D surface models of ground truth and automatic segmentation for mandible #2 of public mandible test dataset: black wireframe, ground truth (operator A); solid green, automatic segmentation result. (a) right lateral view; (b) upper view.

4.4.3. Training and automatic segmentation times

Training time for one fold on one GPU was about 48 hours (1,000 epochs). The trained model provided automatic segmentation of 1 CT scan in approximately 10 minutes.

4.5. Discussion

The main goal of this study was to evaluate the performance of the nnU-Net framework for semantic segmentation of CMF CT scans obtained for the planning of orthognathic surgeries. We chose this deep learning framework because it is open source and well-maintained, and has the ability to automatically configure itself without manual intervention. It has delivered state-of-the art segmentation results on a large diversity of biomedical datasets, surpassing most highly-specialized algorithms (Isensee et al. 2021). Our quantitative results demonstrated the model's relevance for CT scan segmentation, with 97% of our test dataset showing a mean sDSC above 95%. However, this study also illustrated the challenges arising from the evaluation of deep learning-based algorithms in the case of demanding, specific clinical applications which forbid blind trust in quantitative metrics (Nikolov et al. 2021; Reinke et al. 2021 Apr 13). For example, our sDSC results for upper skull masks showed a relatively large dispersion, which is not reflected, however, in expert evaluation results. Indeed, discrepancies between ground truth and prediction masks were mainly found in small bony structures located inside

the skull, which are not relevant for most clinical applications (Figure 34d). Conversely, the number of non-validated teeth segmentation masks cannot be directly correlated to quantitative results, most masks obtaining excellent vDSC and sDSC results (>95%). No teeth labels were rejected based on segmentation errors localized at crown level, despite the difficulty in delineating the upper and lower teeth when they are in contact or show metallic artifacts (Cui et al. 2019; Zhang et al. 2020). Instead, they were rejected because of a few mislabeled voxels located in zones (root apices) clinically relevant for personalized implant manufacturing. However, many other clinical applications, such as computer-assisted diagnosis or planning not involving personalized implant manufacturing, would not require such precision in the segmentation of dental apices. Finally, the clinical value of sDSC metrics in the context of small object segmentation was demonstrated by our mandibular canal segmentation results, for which vDSC did not seem like an appropriate metric (Nikolov et al. 2021; Reinke et al. 2021 Apr 13).

This study, which followed best practice guidelines (Schwendicke, Singh, et al. 2021), is the first to train and test a deep learning segmentation model on such a large number of high-resolution CMF CT scans. Our results are comparable or superior to those of previously published studies, despite the more challenging task we faced: all previous results were based on smaller test datasets (between 0 and 30 scans) and fewer segmentation masks (Cui et al. 2019; Minnema et al. 2019; Qiu et al. 2019; Torosdagli et al. 2019; Chung et al. 2020; Jaskari et al. 2020; Kwak et al. 2020; Lian et al. 2020; Zhang et al. 2020; Murabito et al. 2021; Qiu et al. 2021; Wang et al. 2021). No previous work had included a cohort of consecutive patients, and only one previous publication had clearly stated that its database included patients with syndromic conditions (Torosdagli et al. 2019). Our results for the segmentation of scans from patients with marked syndromic deformities show the versatility of the model (Figure 34b), and are comparable to those of Wang *et al.* who recently reported on a deep learning-based model for multi-task segmentation of bone and teeth separately (Wang et al. 2021). However, the latter study lacked a hold-out test dataset, included only scans devoid of metal artifacts and did not differentiate between maxillary and mandibular structures. Our results for the segmentation of mandibular canals (vDSC of 81.59% ± 5.79%, sDSC of 97.9% ± 3.51%) are significantly superior to those reported by Jaskari *et al.* (vDSC of 57%) and Kwak *et al.* (average Jaccard coefficient of canal and background of 57.72%) (Jaskari et al. 2020; Kwak et al. 2020). However, most existing studies used CBCTs, which are more difficult to segment than CT scans (Torosdagli et al. 2019). In addition, the absence of a public dataset of full high-resolution CMF (CB)CT scans with ground truth segmentations prevents direct comparison of our trained nnU-Net model with previously described ones.

This research has several limitations. It is a single-center study, and thus cannot assess the reliability of the results in other cohorts. Generalization of our results to other CT machines or deformities was partially evaluated on 10 public CT scans obtained using a different CT machine and presenting very different anatomies from those in our training dataset. No subjects from our training database had such edentulous mandibles or endotracheal tube. Setting aside mandible #10 on the basis of endotracheal tube interference, our vDSC results nonetheless outperformed semi-automatic open source segmentation algorithms tested on the same dataset (Wallner, Schwaiger, et al. 2019). This suggests that the trained nnU-Net model presented in this study could be used in other settings and for other clinical applications in the fields of CMF or dentistry requiring high-resolution CT scans, although proper multicenter and prospective evaluations are still needed. In order to assess the efficacy of this trained nnU-Net model in clinical practice, we plan on evaluating its use prospectively for orthognathic surgery planning and personalized implant manufacturing. As our deep learning model was trained with a target size of $0.31*0.45*0.45mm^3$, it is not expected to provide competitive results on CT scans with much lower resolutions (resolutions around $2.5*1*1mm^3$ used in other clinical contexts (Nikolov et al. 2021), for example). Our study focused on CT scans because it is the imaging modality we use for computer-assisted planning of orthognathic surgery at this time. In future works we plan to fine-tune this model in order to evaluate its performance with CBCT scans, another widespread and challenging imaging modality. Finally, we will attempt to use our segmentation results for automatic localization of anatomic landmarks, in order to provide cephalometric measurements for clinical diagnosis and treatment planning (Torosdagli et al. 2019; Dot et al. 2020; Lian et al. 2020; Wang et al. 2021).

To conclude, this study showed that nnU-Net could be used out-of-the-box (along with a simple post-processing volume filter) to provide robust segmentation of routine preoperative CMF CT scans. While the successful segmentation of dental apices will require additional efforts, quantitative results and industry expert evaluation demonstrated the clinical validity of our trained model for the segmentation phase of computer-assisted orthognathic surgery planning. Our results suggest that the nnU-Net framework could be trained from scratch easily, using databases from other departments, to answer the specific needs of many clinical setting.

5 : Évaluation du potentiel de généralisation du modèle

5.1. Objectifs

Le modèle de segmentation présenté dans le Chapitre 4 présente des résultats très prometteurs sur un jeu de données de test constitué de 153 CT-Scan. Cependant, 94.1% (*n* = 144) de ces imageries provenaient du même service hospitalier et de la même machine, ce qui peut questionner le potentiel de généralisation de nos résultats dans d'autres situations cliniques.

L'objectif de cette étude complémentaire a été d'évaluer le modèle de segmentation présenté au Chapitre 4 sur une base de données d'imageries CT-Scans externes à notre service hospitalier.

5.2. Matériels et Méthodes

5.2.1. Sélection des patients

Un partenaire industriel, la société Materialise, a sélectionné aléatoirement 25 imageries CT-Scans qui lui ont été transmises dans le cadre d'une planification de chirurgie orthognathique. Les accords nécessaires à une réutilisation de ces données à des fins de recherche ont été obtenus préalablement à cette étude. Ces imageries ont été effectuées sur diverses machines CT-Scans, leurs caractéristiques principales sont détaillées dans le Tableau 11.

Tableau 11 : Caractéristiques des CT Scans dans le jeu de données de test Externe.

	External test
Number of CT scans	25
Mean in-plane pixel size (mm^2)	0.41 * 0.41
Mean slice thickness (mm)	0.67
Number of scans by CT Machine, no. (%)	
GEHC Discovery CT750 HD	3 (12)
GEHC Revolution EVO	5 (20)
GEHC Discovery MI	2 (8)
Siemens SOMATOM Definition Flash	8 (32)
Siemens SOMATOM Force	2 (8)
Other CT Machine[1]	5 (20)

GEHC: GE Healthcare. [1]Toshiba Aquilion PRIME, Philips Ingenuity CT, Siemens Perspective, GEHC Revolution CT.

5.2.2. Segmentation de référence

La segmentation de référence (« *ground truth* ») des imageries a été effectuée selon le procédé industriel présenté au paragraphe 4.3.2. Quatre masques de segmentation ont été obtenus : massif facial moyen et supérieur (« *upper skull* »), mandibule (« *mandible* »), dents supérieures (« *upper teeth* ») et dents inférieurs (« *lower teeth* »). Le canal mandibulaire n'a pas été segmenté sur ces données. Ces données n'ont pas été partagées avec nous et sont restées chez le partenaire industriel.

5.2.3. Segmentation automatisée

Nous avons effectué la segmentation des 25 imageries CT-Scan en suivant la méthode présentée au Chapitre 4. Le processus d'inférence a été effectué une seule fois, et les résultats (masques de segmentation) ont été transmis au partenaire industriel pour l'évaluation des résultats.

5.2.4. Évaluation des résultats

Le partenaire industriel s'est chargé d'effectuer l'évaluation quantitative des résultats. Les mesures utilisées ont été le DSC volumétrique (vDSC) et le DSC surfacique à 1 mm (sDSC). Nous avons effectué une évaluation qualitative complémentaire, en analysant les modèles 3D reconstruits à partir des masques de segmentation.

5.3. Résultats

5.3.1. Caractéristiques des patients

Le Tableau 12 présente les caractéristiques principales des patients de cette base de données externe. Tous les patients présentaient des dents maxillaires et mandibulaires.

Tableau 12 : Caractéristiques des patients du jeu de données de test externe.

Characteristic	External test (*n = 25)*
Age, mean ± SD, years	25.6 ± 8.7
Gender, no. (%)	
Female	12 (48)
Male	13 (52)
Skeletal deformity, no. (%)	
Class I	1 (4)
Class II[a]	12 (48)
Class III[b]	12 (48)
Metal artifacts, no. (%)	
Orthodontic materials	17 (68)
Metallic dental filling/crown	6 (24)
Fixation implants[c]	1 (4)
No metallic artifact	7 (28)

[a]Prognathic maxilla and/or retrognathic mandible. [b]Retrognathic maxilla and/or prognathic mandible. [c]Fixation implants from a previous surgery. SD, standard deviation

5.3.2. Évaluation quantitative

Les moyennes ± écarts types des vDSC et sDSC pour les 4 masques était respectivement de 94.07 ± 1.96% et 98.81 ± 0.7%. Les résultats moyens des vDSC et sDSC pour chaque masque de segmentation sont présentés dans le Tableau 13, et la Figure 38 montre la distribution des résultats. L'ensemble des masques de segmentation présentaient un sDSC supérieur à la limite de significativité clinique fixée à 95%.

Tableau 13 : Résultats moyens de vDSC et sDSC sur le jeu de données de test externe (*n = 25*).

Metric, mean ± SD [%]	Upper Skull	Mandible	Upper teeth	Lower Teeth	Total
vDSC	92.64 ± 2.11	94.07 ± 1.44	95 ± 1.63	94.58 ± 1.82	94.07 ± 1.96
sDSC	98.59 ± 0.75	98.72 ± 0.72	99.01 ± 0.6	98.94 ± 0.66	98.81 ± 0.7

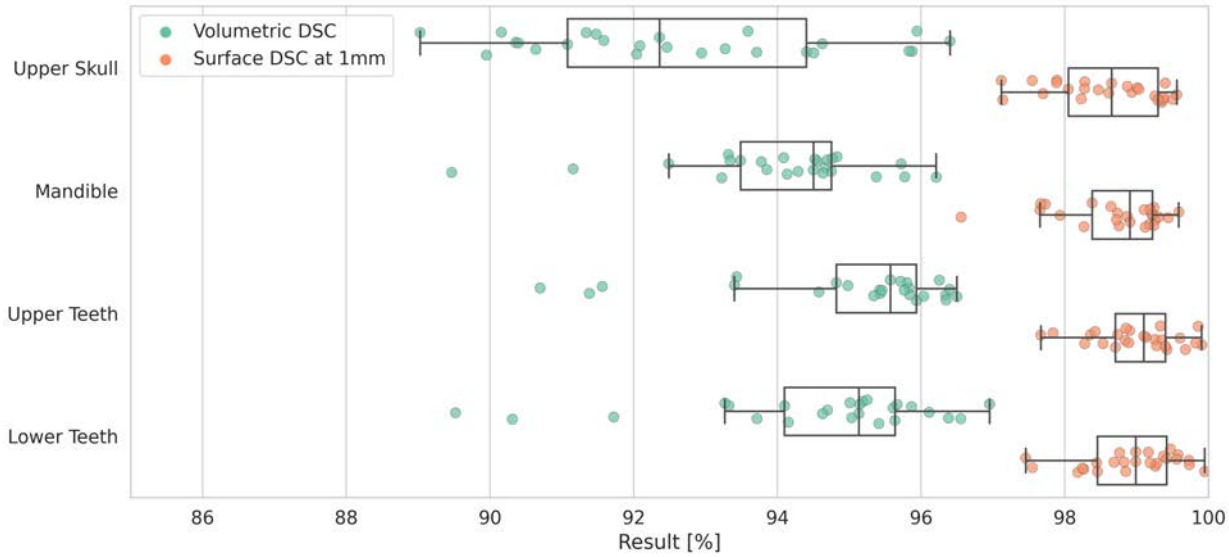


Figure 38 : Distribution des résultats de vDSC et sDSC sur le jeu de données de test externe (*n = 25*), pour chaque masque de segmentation. Aucun résultat n'est inférieur à 85%.

5.3.3. Évaluation qualitative

L'examen des modèles 3D reconstruits à partir des masques de segmentation montre chez plusieurs sujets des défauts de segmentation de zones du massif facial moyen et supérieur présentant de faibles épaisseurs osseuses (Figure 39). Ces défauts semblent plus présents que dans nos résultats obtenus au Chapitre 4.

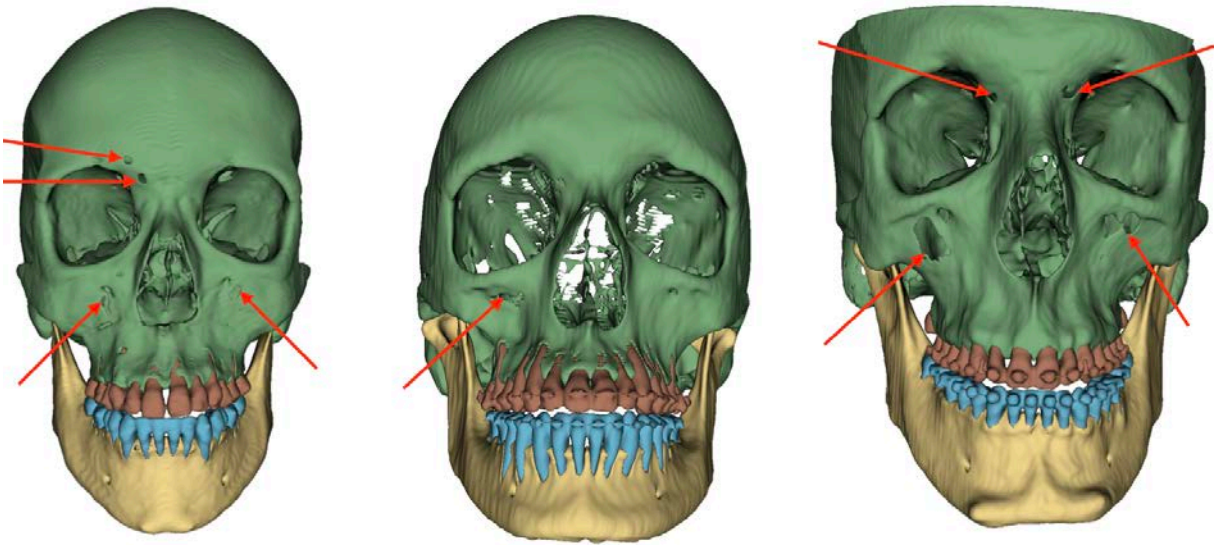


Figure 39 : Défauts de segmentation (flèches rouges) au niveau de la segmentation du massif facial moyen et supérieur sur les modèles 3D de 3 sujets.

5.4. Discussion

Le modèle de segmentation présenté dans le Chapitre 4 a correctement segmenté les 25 imageries CT-Scan issues d'une base de données externe à notre service hospitalier. Ces imageries provenaient pour leur très grande majorité (88 %) d'une autre machine que la GEHC Discovery CT750 HD qui était largement majoritaire dans notre précédente base de données de test. Les résultats quantitatifs sont comparables voire meilleurs que ceux que nous présentions au Chapitre 4, aucun masque de segmentation n'ayant un sDSC inférieur à 95 %. A noter que les sujets présentaient moins d'artéfacts métalliques que ceux de notre base de données de test présentée au Chapitre 4.

Les défauts retrouvés au niveau de certaines zones du massif facial moyen et supérieur ont également été décrits par d'autres auteurs (Liu et al. 2021). Si la segmentation de ces zones est nécessaire pour une utilisation clinique à venir, ces défauts pourraient être comblés sur les modèles surfaciques 3D par une méthode d'interpolation locale.

Ces résultats montrent le potentiel de généralisation et d'utilisation clinique de notre modèle. Une évaluation clinique prospective et multicentrique, incluant un plus grand nombre d'imageries, reste nécessaire pour déterminer l'efficacité et les limites du modèle dans un contexte de soin.

PARTIE C :

Localisation des points céphalométriques

Résumé

Cette troisième partie s'intéresse à la localisation des points céphalométriques 3D sur des imageries dento-maxillo-faciales.

Le Chapitre 6 présente une étude de répétabilité et reproductibilité du placement manuel des points céphalométriques 3D sur des imageries CT-Scan. Trois opérateurs ont localisé à deux répétitions 33 points céphalométriques sur 20 imageries CT-Scan issues de notre base de données d'imageries pré-chirurgicales. Les 33 points céphalométriques ont été sélectionnés pour inclure à la fois des points « traditionnels », adaptations 3D des points précédemment utilisés en 2D, et des points « nouveaux » qui ne pouvaient pas être localisés sur les imageries 2D.

Les résultats de cette étude permettent d'établir une référence concernant la fiabilité du placement de ces points céphalométriques 3D sur des imageries pré-chirurgicales, ce qui n'avait jamais été fait pour plusieurs de ces « nouveaux » points. Ce travail a été publié en 2021.

Le Chapitre 7 s'intéresse à l'automatisation de la localisation des points céphalométriques 3D sur des imageries CT-Scan. Nous avons annoté les 33 points céphalométriques décrits précédemment sur 198 des imageries pré-chirurgicales de notre base de données. Ces données ont été utilisées pour entraîner et évaluer une approche d'automatisation reposant sur l'enchaînement de plusieurs réseaux d'apprentissage profond présentant l'architecture Spatial Configuration-Net.

Cette approche originale permet de localiser entièrement automatiquement les 33 points en une minute. L'évaluation des résultats sur une base de données de test de 38 imageries CT-Scan est très encourageante, avec une erreur moyenne pour l'ensemble des points de 1.0 ± 1.3mm et 90 % des points situés à moins de 2 mm de la référence manuelle. Pour les points osseux, ces résultats sont équivalents à ceux de notre étude de répétabilité et reproductibilité du placement manuel des points et pourraient être utilisés pour effectuer des mesures céphalométriques. Pour les points dentaires, des améliorations sont encore nécessaires afin d'obtenir des mesures céphalométriques plus fiables.

Ces résultats sont comparables aux meilleurs travaux publiés dans la littérature, qui n'étaient pas évalués sur des bases de données de test mais uniquement suivant des méthodes de validations croisées. Cette étude a été publiée en 2022.

6 : Reproductibilité du placement manuel des points céphalométriques 3D

Ce chapitre a fait l'objet d'une publication dans le *Journal of Clinical Medicine* en 2021, sous le titre :

Three-Dimensional Cephalometric Landmarking and Frankfort Horizontal Plane Construction: Reproducibility of Conventional and Novel Landmarks

Gauthier Dot[1,2], Frédéric Rafflenbeul[3], Adeline Kerbrat[1], Philippe Rouch[1], Laurent Gajny[1], Thomas Schouman[1,4]

[1] Institut de Biomecanique Humaine Georges Charpak, Arts et Metiers Paristech, Paris, France ;

[2] Universite de Paris, AP-HP, Hopital Pitie-Salpetriere, Service d'Odontologie, Paris, France ;

[3] Department of Dento-Facial Orthopedics, Faculty of Dental Surgery, Strasbourg University, Strasbourg, France;

[4] Sorbonne Universite, AP-HP, Hopital Pitie-Salpetriere, Service de Chirurgie Maxillo-Faciale, Paris, France.

DOI : https://doi.org/10.3390/jcm10225303

6.1. Abstract

In some dentofacial deformity patients, especially patients undergoing surgical orthodontic treatments, Computed Tomography (CT) scans are useful to assess complex asymmetry or to plan orthognathic surgery. This assessment would be made easier for orthodontists and surgeons with a three-dimensional (3D) cephalometric analysis, which would require the localization of landmarks and the construction of reference planes. The objectives of this study were to assess manual landmarking repeatability and reproducibility (R&R) of a set of 3D landmarks and to evaluate R&R of vertical cephalometric measurements using two Frankfort Horizontal (FH) planes as references for horizontal 3D imaging reorientation. Thirty-three landmarks, divided into "conventional", "foraminal" and "dental", were manually located twice by 3 experienced operators on 20 randomly-selected CT scans of orthognathic surgery patients. R&R confidence intervals (CI) of each landmark in the -x, -y and -z directions were computed according to the ISO 5725 standard. These landmarks were then used to construct 2 FH planes: a conventional FH plane (orbitale left, porion right and left) and a newly

proposed FH plane (midinternal acoustic foramen, orbitale right and left). R&R of vertical cephalometric measurements were computed using these 2 FH planes as horizontal references for CT reorientation. Landmarks showing a 95% CI of repeatability and/or reproducibility > 2mm were found exclusively in the "conventional" landmarks group. Vertical measurements showed excellent R&R (95% CI < 1mm) with either FH plane as horizontal reference. However, the 2 FH planes were not found to be parallel (absolute angular difference of 2.41°, SD 1.27°). Overall, "dental" and "foraminal" landmarks were more reliable than the "conventional" landmarks. Despite the poor reliability of the landmarks orbitale and porion, the construction of the conventional FH plane provided a reliable horizontal reference for 3D craniofacial CT scan reorientation.

6.2. Introduction

Diagnosis and planning of orthodontic and maxillofacial treatments rely heavily on X-ray imaging. Two-dimensional (2D) X-rays are routinely used but result in a flattening of three-dimensional (3D) craniofacial structures. In some clinical cases of dentofacial deformities – especially patients undergoing surgical orthodontic treatments (orthognathic surgery) – Computed Tomography (CT) or Cone Beam CT (CBCT) scans are useful (Kapila and Nervina 2015). For example, 3D imaging makes it possible to assess complex asymmetry and to obtain highly accurate orthognathic surgery planning that can subsequently be used for the manufacturing of surgical guides (Patel et al. 2011; Kapila and Nervina 2015; Schouman et al. 2015; Pietzka et al. 2019; Quast et al. 2019). Several methods have recently been proposed for a fully automatic detection of the best symmetry plane in craniofacial CT scans (Di Angelo et al. 2019; Noori et al. 2020). The diagnostic value of these scans would increase if they could be used to perform 3D cephalometric analysis, which would require the localization of landmarks (Pittayapat et al. 2014). At the time being, however, no set of 3D landmarks has been deemed sufficiently reproducible and repeatable for 3D cephalometry (Pittayapat et al. 2014; Sam et al. 2018).

Most of the time, three-dimensional cephalometric landmarks previously tested in repeatability and reproducibility studies derived from classic 2D analysis (Sam et al. 2018). Some of these landmarks have been shown to be poorly reproducible in 3D, especially orbitale (Or), porion (Po), gonion (Go), condylion (Co) and ramus (Ra) (Chien et al. 2009; de Oliveira et al. 2009; Lagravère et al. 2010; Olszewski et al. 2010; Schlicher et al. 2012; Titiz et al. 2012; Hassan et al. 2013; Naji et al. 2014; da Neiva et al. 2015; Zamora et al. 2012). The localization of midsagittal landmarks has generally been shown to be reliable, mostly in datasets of patients showing no asymmetries (Pittayapat et al. 2014; Sam et al. 2018). Several authors suggested using "new" landmarks which cannot be localized on 2D

X-rays. More specifically, landmarks located on the craniofacial foramens are presumably easy to identify and should provide good reproducibility (Lagravère et al. 2010; Naji et al. 2014; Sam et al. 2018). However, few studies have tested the reproducibility of the new landmarks, and their reliability has not been tested yet in the context of presurgical orthodontic patients (Naji et al. 2014; Pittayapat et al. 2018).

The main goal of cephalometric landmarking is to measure distances and angles between landmarks and planes so as to obtain a cephalometric analysis. In order to provide clinically relevant measurements that can be decomposed in the 3 planes of space (i.e. anteroposterior, vertical and transversal), 3D images need to be reoriented in a generic coordinate system (Ruellas et al. 2016; Shahen et al. 2018). The Frankfort Horizontal (FH) plane, used for standardizing and unifying the measurements, is the most commonly used horizontal reference for this coordinate system (Ruellas et al. 2016; Santos et al. 2017). Its 3D clinical value has been demonstrated for assessing craniofacial morphology and evaluating soft-tissue and skeletal cants in patients receiving orthognathic surgery (Oh et al. 2013; Lin et al. 2015; Lonic et al. 2017). This plane is conventionally defined in 3D by the 3 following points: left orbitale (Or-L), right porion (Po-R) and left porion (Po-L) (Santos et al. 2017). Hence, this reference plane is based on landmarks that are known to be poorly reproducible in 3D, suggesting that the conventionally defined FH plane is poorly reproducible (Pittayapat et al. 2018). However, landmark reproducibility does not necessarily result in plane reproducibility, as the latter depends on the direction of landmark errors (Lagravère et al. 2010). To our knowledge, no study has yet tested the repeatability and reproducibility of vertical cephalometric measurements using the conventional FH plane (constructed from 3 landmarks) as a reference for horizontal head reorientation.

Looking for a new plane which would remain parallel with the conventional FH plane but be based on more reliable landmarks, Pittayapat *et al*. suggested a novel FH plane, in which the internal acoustic foramina (IAF) would replace Po (Pittayapat et al. 2018). Results from experiments performed on CBCT scans of dry human skulls revealed that the localization of IAF provided better reproducibility than that of Po. Moreover, the authors suggested that another new FH plane, based on mid-IAF, Or-R and Or-L, might replace the conventional FH plane, the angular difference found between the two planes being inferior to 1 degree. These results have not been validated yet on 3D scans of living human subjects.

In this context, using a dataset of preoperative CT scans, the aims of our study were:

1- to assess landmarking repeatability and reproducibility of a set of 33 landmarks containing “conventional”, “foraminal” and “dental” landmarks;
2- to assess repeatability and reproducibility of vertical cephalometric measurements using either the conventional or the newly proposed FH planes as references for horizontal head reorientation;
3- to assess the parallelism between the conventional and the newly proposed FH planes.

6.3. Subjects and Methods

6.3.1. Dataset

Sample size calculation was performed in order to ensure an uncertainty in the repeatability or reproducibility result of 15% for 6 repetitions (McAlinden et al. 2015). As a result, a sample of at least 17 subjects was needed for this study. We performed a random selection of 20 CT scans (7 males, 13 females, mean age 25 ± 8 years) in a database of 134 consecutive orthognathic surgery patients (49 males, 85 females, mean age 27 ± 10 years) from a single Maxillofacial Surgery Department. Patients were considered for inclusion whatever maxillomandibular deformity they presented, with no minimum age. Exclusion criteria were refusal to participate in the research (all patients were contacted by mail) and lack of CT scan segmentation. We used a random number generator to obtain a random sequence of 20 numbers, which was used to select the sample of CT scans included in this study. Allocation was performed by one operator (#1) and supervised by a second operator (#2) at the beginning of the study. All selected subjects showed marked skeletal deformities: 14 skeletal class II - prognathic maxilla and/or retrognathic mandible - (10 short faces, 4 long faces) and 6 skeletal class III - retrognathic maxilla and/or prognathic mandible - (2 short faces, 4 long faces). Six subjects exhibited mandibular asymmetry (2 severe, 4 slight) and 2 subjects exhibited syndromic or rare dentofacial deformities (cleidocranial dysplasia and oligodontia, respectively). A set of 5 random CT scans not included in this study was used for operator training prior to landmarking.

The 20 CT scans were acquired on a Discovery CT750 HD scanner (GE Healthcare, Chicago, USA) set at 100kVp, 50mAs, exposure time 730ms, slice thickness 0.625mm and slice increment 0.320mm. Field of view ranged from 200 to 267mm and pixel size ranged from 0.39 to 0.52mm. Scans were not reoriented after their acquisition. Segmentation of the bones (upper skull, mandible) and upper/lower teeth was performed prior to the study according to an industry-certified semi-automatic process (Materialise, Leuven, Belgium). The study was approved by an Institutional Review Board (IRB No. CRM-2001-051), and all experiments were performed in accordance with relevant guidelines and

regulations. Informed consent was obtained from all subjects and/or their legal guardian(s) for participants below age 16 years (all patients were contacted by mail).

6.3.2. Landmark annotation

The 33 landmarks (Figure 40) were divided into 3 groups: “conventional” (Table 14) “foraminal” (Table 15) and “dental” (Table 16). Operators #1, #2 and #3 (2 trained orthodontists with at least 5 years of clinical experience, 1 final year postgraduate maxillofacial surgeon) received written and verbal instructions on the 3D description and annotation procedure for each landmark (Supplementary Materials C1). Manual reorientation of the CT scans was performed based on the Frankfort Horizontal plane construction obtained from the annotation process. A calibration session was organized before the study began, and the instructions were repeated to the operators once more before the second annotation session.

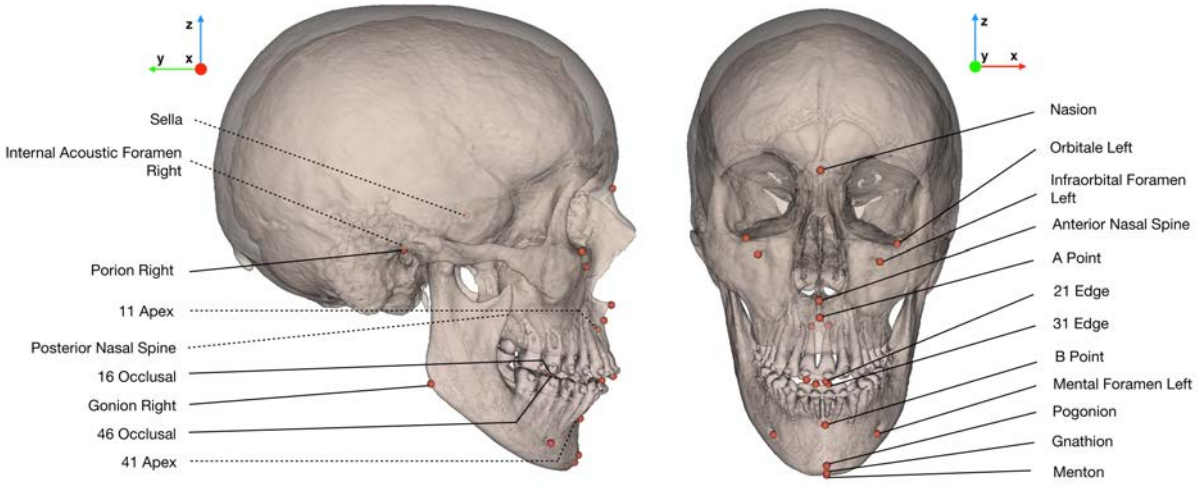


Figure 40: Illustration of the set of 33 landmarks localized by the operators, and the new coordinate system used for statistical analysis. In the case of bilateral landmarks, only one of the two landmarks is labelled. Dotted lines show landmarks localized inside bony structures.

The 20 CT scans and their segmentations were handed over to the 3 operators without any annotations. Manual placement of the 33 landmarks was performed independently by the operators on the software Mimics (v.22.0, Materialise, Leuven, Belgium), and was repeated once after a 3-week interval. Landmarks could be annotated either on the 3D surface or in the Multi-Planar Reconstruction (MPR) views. The operators had neither access to each other’s results nor to their first session’s results when performing the second session. For each session and each CT scan, results were exported as an .xml file containing the x-, y-, z- coordinates of each landmark. The time needed for each CT scan annotation was recorded by the operators and exported in a spreadsheet.

Table 14: Definition of "conventional" landmarks localized in our study (L/R: Left/Right).

Landmark name	Description
Nasion (Na)	Medial (and upper) point of the frontonasal suture
Orbitale L/R (Or-L / Or-R)	Lowest point of the orbital rim L/R
Anterior Nasal Spine (ANS)	Medial and most anterior point of the nasal spine
A Point (A)	Medial and most posterior point of the maxilla
B Point (B)	Medial and most posterior point of the mandible
Pogonion (Pog)	Medial and most anterior point of the mandible
Gnathion (Gn)	Medial and midpoint between Pog and Me
Menton (Me)	Medial and lowest point of the mandible
Gonion L/R (Go-L / Go-R)	Midpoint of the gonial angle L/R
Porion L/R (Po-L / Po-R)	External & uppermost point of the auditory canal L/R
Posterior Nasal Spine (PNS)	Medial & most distal point of the osseous palate
Sella (S)	Central point of the sella

Table 15: Definition of "foraminal" landmarks localized in our study (L/R: Left/Right).

Landmark name	Description
Infraorbital Foramen L/R (IF-L / IF-R)	External & most distal point of the infraorbital foramen L/R
Mental Foramen L/R (MF-L / MF-R)	External & most mesial point of the mental foramen L/R
Internal Acoustic Foramen L/R (IAF-L / IAF-R)	External, most mesial and posterior point of the internal acoustic foramen L/R

Table 16: Definition of "dental" landmarks localized in our study (FDI World Dental Federation notation for teeth numbering).

Landmark name	Description
11, 21, 31, 41 edges (11E, 21E, 31E, 41E)	Midpoint of 11/21/31/41 incisal edges
11, 21, 31, 41 apexes (11A, 21A, 31A, 41A)	Root apex of 11/21/31/41
16, 26 occlusal (16O, 26O)	Summit of the mesiopalatal cusp of 16/26
36, 46 occlusal (36O, 46O)	Central fossa of 36/46

6.3.3. Statistical analysis

6.3.3.1. New coordinate system for each CT scan

After the two annotation sessions, each CT scan was reoriented in a new coordinate system according to the mean Frankfort Horizontal plane resulting from the 6 repetitions (mean Po-R, mean Po-L, mean Or-L). The origin was set at mid-porion; the -x axis followed the sagittal plane (from right to left); the -y axis followed the frontal plane (from front to back); and the -z axis followed the axial plane (from toe to head) (Figure 1). The landmarking results were then referenced in the new coordinate system before performing the statistical analysis.

6.3.3.2. Landmark repeatability and reproducibility

For each landmark, repeatability and reproducibility standard deviations (SD) were computed according to the ISO 5725 standard of the International Organization for Standardization (ISO 5725-2:2019). Upon initial inspection of the results, the standard's recommendations were followed for clear outlier points, whose annotations were considered as missing data. The reliability of each landmark in the -x, -y and -z directions was then estimated, considering a 95% confidence interval (CI) of 2*SD of repeatability and reproducibility. Modified Bland-Altman plots, showing the deviations of the landmark positions from their means for the 20 CT scans, were computed for each landmark and direction (Donatelli and Lee 2013a; Donatelli and Lee 2013b).

6.3.3.3. Repeatability and reproducibility of vertical measurements with the conventional FH plane and the newly proposed FH plane

For each CT scan and landmarking session (3 operators, 2 repetitions), we computed the landmarks' orthogonal projections on 2 FH planes: the conventional FH plane (Or-L, Po-R, Po-L) and the newly proposed FH plane (Or-R, Or-L, mid-IAF). The results were used to compute the standard deviations of repeatability and reproducibility (ISO 5725 standard) of the landmarks' vertical measurements, using the 2 FH planes as horizontal reference.

6.3.3.4. Parallelism between conventional and newly proposed FH planes

In order to assess whether the conventional FH plane and the new FH plane were parallel, the orthogonal projections of points IAF-R and IAF-L were computed on the mean conventional FH plane (as defined previously) for each subject. We then computed the absolute angular differences between

the conventional FH plane and the novel FH plane, using trigonometry to calculate the angles between the normals to the planes.

6.3.3.5. Time needed for landmark localization

Mean time needed and standard deviation for landmark localization were computed.

All data were analysed using the softwares Matlab (v.R2020a, MathWorks, Natick, MA, USA) and RStudio (v.1.3, RStudio PBC, Boston, MA, USA).

6.4. Results

6.4.1. Landmark repeatability and reproducibility

Outliers were identified for mental foramen points right/left placed by operator #3 during the first annotation session (subjects 4 to 20) and were considered as missing data (Supplementary Materials C2). Repeatability and reproducibility results for the 33 landmarks are shown in Table 17. The landmarks with 95% CI of repeatability and/or reproducibility superior to 2mm for one of their axes were exclusively found in the “conventional” landmark group: point B (-z axis), gonion right/left (-y and –z axes), orbitale right/left (-x axis) and porion right/left (-x axis). Figure 41 shows an example of the modified Bland-Altman plots obtained for five left landmarks: three “foraminal” landmarks (IAF-L, infraorbital foramen left (IF-L), mental foramen left (MF-L)) and two “conventional” landmarks (Or-L and Po-L).

Table 17: 95% confidence interval (2*SD) of repeatability and reproducibility of the landmarks (mm), following the ISO 5725 standard. Values between 1 and 2mm are highlighted in orange, and values superior to 2mm are highlighted in red. Repet., repeatability; Repro., reproducibility; SD, standard deviation; L/R: Left/Right.

Landmark	X Axis		Y Axis		Z Axis	
	Repet. 2*SD	Repro. 2*SD	Repet. 2*SD	Repro. 2*SD	Repet. 2*SD	Repro. 2*SD
11 Apex	0.24	0.35	0.38	0.58	0.28	0.38
11 Edge	0.25	0.34	0.24	0.30	0.08	0.10
16 Occlusal	0.63	0.76	1.27	1.51	0.21	0.28
21 Apex	0.30	0.41	0.33	0.58	0.28	0.47
21 Edge	0.29	0.39	0.26	0.34	0.05	0.08
26 Occlusal	0.65	0.83	0.68	0.88	0.17	0.23
31 Apex	0.18	0.24	0.25	0.32	0.27	0.36
31 Edge	0.40	0.26	0.18	0.25	0.14	0.10
36 Occlusal	0.63	0.87	1.15	1.43	0.38	0.47
41 Apex	0.18	0.22	0.35	0.50	0.23	0.35
41 Edge	0.19	0.24	0.20	0.22	0.05	0.07
46 Occlusal	0.40	0.54	0.55	0.82	0.23	0.30
A Point	0.80	0.86	0.29	0.34	1.31	1.59
Anterior Nasal Spine	0.57	0.67	0.80	1.31	0.57	1.10
B Point	0.65	1.12	0.55	0.63	2.46	2.89
Gnathion	0.75	1.14	0.76	0.92	1.05	1.25
Gonion L	0.63	0.93	1.55	2.02	1.96	2.64
Gonion R	0.61	0.82	1.38	1.75	1.94	2.45
Infraorbital Foramen L	0.88	0.93	0.81	1.01	0.63	0.93
Infraorbital Foramen R	0.87	0.99	0.80	0.93	0.66	0.82
Internal Acoustic Foramen L	0.52	0.79	0.81	1.09	0.84	1.15
Internal Acoustic Foramen R	0.51	0.82	0.88	1.04	0.56	1.01
Mental Foramen L	0.23	0.38	0.26	0.47	0.30	0.47
Mental Foramen R	0.23	0.35	0.25	0.45	0.32	0.43
Menton	0.76	1.11	1.29	1.84	0.37	0.54
Nasion	0.41	0.52	0.22	0.27	0.62	0.84
Orbitale L	2.05	3.40	1.00	1.82	0.40	0.56
Orbitale R	1.83	3.23	1.07	1.75	0.37	0.46
Posterior Nasal Spine	0.29	0.53	0.36	0.50	0.61	0.92
Pogonion	0.71	1.15	0.38	0.49	1.64	1.99
Porion L	2.39	2.84	0.88	1.28	0.61	0.71
Porion R	2.16	2.73	1.13	1.37	0.69	0.87
Sella	0.94	1.12	0.60	0.70	0.82	1.15

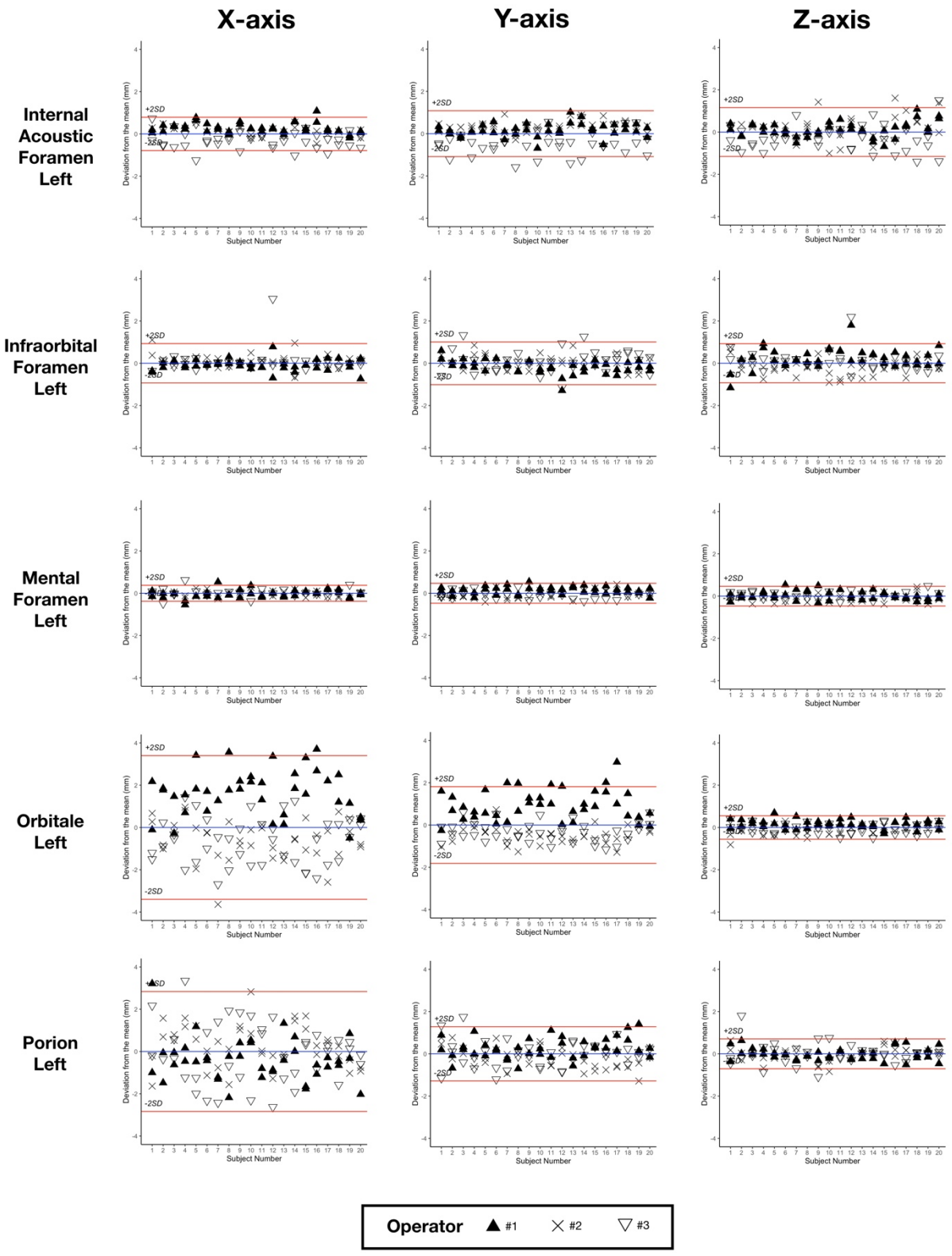


Figure 41: Bland-Altman plots for five left landmarks, showing the deviations from the mean (blue line) of the 6 repetitions for the 20 subjects. Red lines show the ± 2*SD of reproducibility. SD, standard deviation.

6.4.2. Repeatability and reproducibility of conventional and newly proposed FH planes

The results of the repeatability and reproducibility analysis of vertical measurements of the landmarks using the 2 different FH planes as horizontal reference are shown in Table 18.

Table 18: 95% confidence interval (2*SD) of repeatability and reproducibility of the vertical measurements of the landmarks (mm) using 2 FH planes as horizontal references, following the ISO 5725 standard. FH, Frankfort Horizontal plane; Repet., repeatability; Repro., reproducibility; SD, standard deviation; L/R: Left/Right.

Landmark	Conventional FH		Novel FH	
	Repet. 2*SD	Repro. 2*SD	Repet. 2*SD	Repro. 2*SD
11 Apex	0.54	0.74	0.35	0.46
11 Edge	0.61	0.82	0.40	0.54
16 Occlusal	0.51	0.68	0.29	0.39
21 Apex	0.53	0.71	0.35	0.46
21 Edge	0.58	0.78	0.40	0.54
26 Occlusal	0.37	0.54	0.31	0.43
31 Apex	0.52	0.68	0.32	0.42
31 Edge	0.57	0.75	0.37	0.49
36 Occlusal	0.35	0.52	0.31	0.43
41 Apex	0.52	0.69	0.31	0.42
41 Edge	0.58	0.77	0.37	0.49
46 Occlusal	0.47	0.64	0.29	0.40
A Point	0.58	0.77	0.37	0.49
Anterior Nasal Spine	0.61	0.81	0.39	0.53
B Point	0.53	0.70	0.33	0.45
Gnathion	0.53	0.70	0.33	0.45
Gonion L	0.48	0.56	0.52	0.74
Gonion R	0.59	0.76	0.46	0.70
Infraorbital Foramen L	0.41	0.58	0.35	0.48
Infraorbital Foramen R	0.62	0.80	0.33	0.43
Internal Acoustic Foramen L	0.60	0.71	0.61	0.91
Internal Acoustic Foramen R	0.63	0.80	0.59	0.90
Mental Foramen L	0.40	0.55	0.32	0.43
Mental Foramen R	0.55	0.72	0.30	0.39
Menton	0.50	0.67	0.31	0.42
Nasion	0.56	0.75	0.35	0.46
Orbitale L	0.40	0.56	0.38	0.50
Orbitale R	0.65	0.84	0.34	0.43
Posterior Nasal Spine	0.39	0.53	0.31	0.46
Pogonion	0.54	0.71	0.33	0.46
Porion L	0.61	0.71	0.65	0.93
Porion R	0.69	0.87	0.59	0.88
Sella	0.43	0.55	0.40	0.60

6.4.3. Parallelism between conventional and novel FH planes

When using the mean conventional FH plane as horizontal reference, the mean absolute vertical measurements ± SD of IAF-L and IAF-R were 2.68 ± 2.51mm and 2.78 ± 2.29mm, respectively. Measurement results for each subject and each repetition are shown in Figure 42. The absolute angular difference between the conventional and the novel FH planes was 2.41° (SD 1.27°).

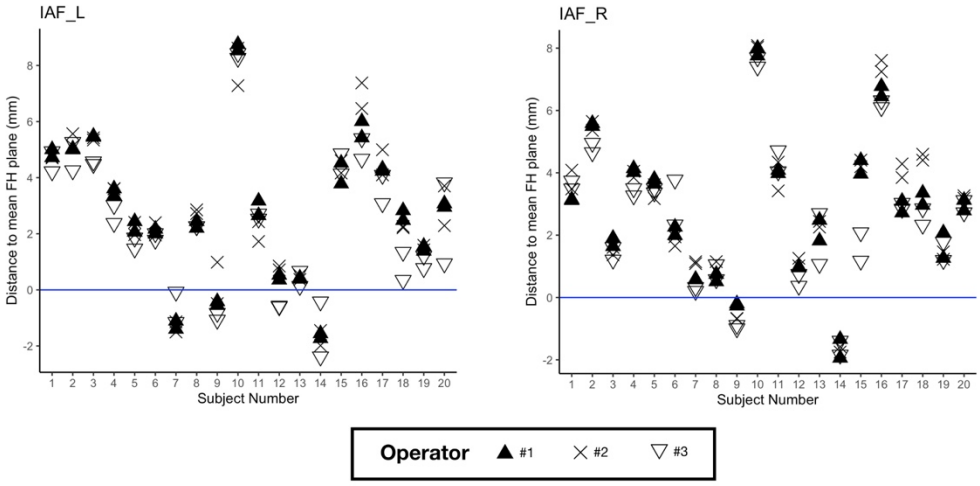


Figure 42: Vertical measurements of IAF left (on the left) and right (on the right) for each subject and repetition, using the mean conventional FH plane as horizontal reference.

6.4.4. Time needed for landmark localization

The average time required to landmark one CT scan was 14:48 ± 03:45 minutes.

6.5. Discussion

The reliability of 3D cephalometric landmarking and Frankfort Horizontal plane construction is a recurrent clinical issue in orthodontics and orthognathic surgery planning. In this study, we performed a repeatability and reproducibility analysis of conventional and 3D-specific cephalometric landmarks using a database of 20 randomly selected routine presurgical CT scans.

The first aim of our study was to assess landmarking reliability in a set of 33 landmarks containing “conventional”, “foraminal” and “dental” landmarks. As in previously published studies, we ranked the landmarks based on the 95% CI results: landmark with clinically acceptable error when the 95% CI was below 1mm; landmark useful in most analyses when the 95% CI was between 1 and 2mm (highlighted in orange in Table 17); landmark to be used with caution when the 95% CI was above 2mm (highlighted

in red in Table 17) (Lagravère et al. 2010; Olszewski et al. 2010). Using this classification, all "dental" and "foraminal" landmarks showed a clinically acceptable error or were considered useful in most analyses (16O, 36O, IF-L, IAF-L, IAF-R). The group of "conventional" landmarks showed several landmarks to use with caution (B point, gonion right and left (Go-R, Go-L), Or-R, Or-L, Po-R, Po-L). These findings are in line with previously published reproducibility studies, in which "conventional" landmarks resulting from 2D cephalometric analysis are subject to caution (Chien et al. 2009; de Oliveira et al. 2009; Lagravère et al. 2010; Olszewski et al. 2010; Schlicher et al. 2012; Titiz et al. 2012; Hassan et al. 2013; Naji et al. 2014; da Neiva et al. 2015; Zamora et al. 2012). As shown in the Bland-Altman plots (Figure 41), "foraminal" landmarks IAF-L, IF-L and MF-L lead to better repeatability and reproducibility than "conventional" landmarks Or-L and Po-L. We chose not to perform statistical tests such as intraclass correlation coefficients or paired t-tests because of their proven inadequacy in measuring landmarking reliability (Donatelli and Lee 2013a; Donatelli and Lee 2013b). Outliers were only found in mental foramen landmarks. The definition of this specific "new" landmark was initially ill-understood by one of the operators, who located the landmark at the distal end of the foramen instead of the mesial end, as was agreed upon (Supplementary Materials C2). This illustrates the challenges encountered with landmark identification, which requires very precise definitions in order to be reliable (Naji et al. 2014). The fact that our dataset was made of presurgical CT scans did not appear to influence the results negatively when compared with non-surgical subjects in the literature (Chien et al. 2009; Lagravère et al. 2010; Olszewski et al. 2010; Schlicher et al. 2012; Titiz et al. 2012; Hassan et al. 2013; Naji et al. 2014; da Neiva et al. 2015).

Our second objective was to assess the repeatability and reproducibility of vertical 3D cephalometric measurements using either the conventional FH plane or the newly proposed FH plane as reference for horizontal head reorientation. The measurements showed excellent repeatability and reproducibility (95% CI<1mm), using either FH plane as reference. These results tend to prove that despite the poor reliability of the landmarks used to construct them, both planes can be used as reliable horizontal references for head reorientation. An explanation could be that the poor reproducibility of Or, Po and IAF points mainly concerns the -x and -y coordinates. Our results show that a simple method using only 3 landmarks can provide a reliable horizontal reference for 3D head reorientation. Other methods, such as manual reorientation of the 3D model along the FH plane (Ruellas et al. 2016) or computation of additional semi-automated landmarks (Shahen et al. 2018) have been shown to be reliable, but are more complex in terms of implementation.

Our third objective was to assess whether the conventional FH plane and the "new" FH plane were parallel. Our results regarding the vertical positions of IAF-L and IAF-R to the conventional FH plane, as well as the angular differences between the 2 planes, show that the planes are not parallel for most subjects. The vertical distance of IAF-L and IAF-R to the conventional FH plane showed significant variations in our cohort (Figure 42). This tends to invalidate the relevance of this "new" FH plane as a replacement for the conventional FH plane. Our results are not consistent with Pittayapat *et al.*'s findings, which reported an angular difference of 0.53 ± 0.37° between the conventional FH plane (called FH1 in their study and constructed using mid-Po, Or-R, Or-L) and their new FH plane. In order to facilitate the comparison between our data and Pittayapat *et al*.'s, the angular distances between the conventional and all newly proposed FH planes are provided in Supplementary Materials C3. The discrepancies between the two studies might be explained by a different definition of the IAF-R and IAF-L points. We tried our best to follow the instructions given by the aforementioned authors to define these points (Supplementary Materials C1), but a more precise anatomical description might be needed.

The average duration of CT scan landmarking confirms that this is a time-consuming procedure, requiring prior training of the operators and potentially limiting clinical implementation. Our durations were in line with Hassan *et al.*'s results, who reported an average time of 10:41 ± 4:01 minutes to localize 22 landmarks using a similar protocol (Hassan et al. 2013). Given the operator training and time needed to place the 3D landmarks manually, semi- or fully automatic landmarking could help advance clinical use of 3D cephalometry (Dot et al. 2020).

This study has three main limitations. Firstly, it was based on a retrospective selection of a limited number of clinical cases, which can be a source of potential bias or imprecision in the statistical analysis. Despite the heterogeneity of our sample (sex, age, skeletal classes), we assumed that the within-subject standard deviation of each landmark was the same for all the patients (Bland and Altman 1996). We believe that our randomly-selected cases from clinical practice are an asset for the clinical applicability of our results because they are representative of the variability encountered in a clinical practice. Secondly, it was performed on CT scans instead of CBCT scans, which are more commonly used for 3D cephalometric studies. We made this choice because CT scanning is the imaging modality currently used for orthognathic surgery planning in our department. We chose to include only presurgical patients in this study because they display a variety of significant craniofacial deformities for which 3D cephalometry could be very beneficial for in-depth evaluation (American Academy of Oral and Maxillofacial Radiology 2013; Kapila and Nervina 2015). Thirdly, the placement of most of our landmarks was carried out on the CT scans' 3D surface models, using the MPR views for

adjustments and verifications. As has been reported previously, while the use of 3D surface models make the annotation process easier and more robust, it implies prior segmentation of the CT scans (Hassan et al. 2013). Performing the segmentation process manually is tedious and time-consuming, but full automatization of the task, an active and promising research field, could resolve this problem (Wang et al. 2021). Given that most of the annotations were made on 3D surface models, we hypothesize that using CBCT scans instead would yield similar results. In order to evaluate the consequences of our results on patients' soft tissues, the 3D surface models used in this study could be superimposed with the patients' facial scans (Granata et al. 2020). The virtual patients obtained using this recently described technique could provide valuable additional clinical insights and help surgical planning. Not having a facial scanner at our disposal, we were not able to test the technique in our study.

Overall, our repeatability and reproducibility study on CT scans showed that "dental" and "foraminal" 3D landmarks tended to be more reliable than "conventional" cephalometric 3D landmarks in presurgical patients. Despite the poor overall reliability of the landmarks orbitale and porion in 3D, the conventional FH plane is a reliable horizontal reference for head reorientation and vertical measurements. The new FH plane, using IAF instead of porion, provided a reliable horizontal reference but was not found to be parallel to the conventional FH plane.

7 : Placement automatisé des points céphalométriques 3D

Ce chapitre a fait l'objet d'une publication dans le *Journal of Dental Research* en 2022, sous le titre :

Automatic 3-Dimensional Cephalometric Landmarking via Deep Learning

Gauthier Dot[1,2], Thomas Schouman[1,3], Shaole Chang[1], Frédéric Rafflenbeul[4], Adeline Kerbrat[1], Philippe Rouch[1,5], Laurent Gajny[1]

[1] Institut de Biomecanique Humaine Georges Charpak, Arts et Metiers Institute of Technology, Paris, France ;

[2] Universite Paris Cite, AP-HP, Hopital Pitie Salpetriere, Service de Medecine Bucco-Dentaire, Paris, France ;

[3] Medecine Sorbonne Universite, AP-HP, Hopital Pitie-Salpetriere, Service de Chirurgie Maxillo-Faciale, Paris, France ;

[4] Department of Dentofacial Orthopedics, Faculty of Dental Surgery, Strasbourg University, Strasbourg, France.

DOI : https://doi.org/10.1177/00220345221112333

7.1. Abstract

The increasing use of three-dimensional (3D) imaging by orthodontists and maxillofacial surgeons to assess complex dentofacial deformities and plan orthognathic surgeries implies a critical need for 3D cephalometric analysis. Although promising methods were suggested to localize 3D landmarks automatically, concerns about robustness and generalizability restrain their clinical use. Consequently, highly trained operators remain needed to perform manual landmarking. In this retrospective diagnostic study, we aimed to train and evaluate a deep learning (DL) pipeline based on SpatialConfiguration-Net for automatic localization of 3D cephalometric landmarks on computed tomography (CT) scans. A retrospective sample of consecutive presurgical CT scans was randomly distributed between a training/validation set (n = 160) and a test set (n = 38). The reference data consisted in 33 landmarks, manually localized once by 1 operator (n = 178) or twice by 3 operators (n = 20, test set only). After inference on the test set, one CT scan showed "very low" confidence level predictions; we excluded it from the overall analysis but still assessed and discussed the corresponding

results. The model performance was evaluated by comparing the predictions with the reference data; the outcome set included localization accuracy, cephalometric measurements and comparison to manual landmarking reproducibility. On the hold-out test set, the mean localization error was 1.0 ± 1.3mm, while success detection rates for 2.0, 2.5 and 3.0mm were 90.4%, 93.6% and 95.4%, respectively. Mean errors were -0.3 ± 1.3° and -0.1 ± 0.7mm for angular and linear measurements, respectively. When compared to manual reproducibility, the measurements were within the Bland-Altman 95% limits of agreement for 91.9% and 71.8% of skeletal and dentoalveolar variables, respectively. To conclude, while our DL method still requires improvement, it provided highly accurate 3D landmark localization on a challenging test set, with a reliability for skeletal evaluation on par with what clinicians obtain.

7.2. Introduction

Three-dimensional (3D) computed tomography (CT) or cone beam CT (CBCT) scans are increasingly used by orthodontists and maxillofacial surgeons for diagnosis and treatment planning purposes. While two-dimensional (2D) radiographs are still sufficient for most of orthodontic patients, 3D scans allow clinicians to assess complex maxillomandibular deformities and craniofacial anomalies, improving diagnosis and treatment planning for those patients (Kapila and Nervina 2015; American Academy of Oral and Maxillofacial Radiology 2013). More specifically, 3D images are now widely used for the planning of computer-assisted orthognathic surgical procedures (Alkhayer et al. 2020). For each patient, this planning is usually performed by a technician, following a surgeon's prescription based on clinical examination and cephalometric analysis of the 3D scans (Xia et al. 2009). Cephalometric analysis is used to measure the deviation of the skeletal and dentoalveolar parts of the maxilla and the mandible in relation to the skull base, using measurements between specific landmarks placed on each of these structures. The reference method for 3D cephalometric analysis is manual landmarking, which requires around 15 minutes for a highly experienced and trained operator (Hassan et al. 2013; Dot et al. 2021).

The automatization of 3D cephalometric landmarking has been an active research field over the last decade, as the clinical dissemination of such a method would decrease the burden of manual landmarking. Two systematic reviews recently reported on the accuracy of such automated methods (Dot et al. 2020; Schwendicke, Chaurasia, et al. 2021). Both yielded promising results for deep learning (DL) based methods, which outperformed previously proposed knowledge-based, atlas-based or shallow learning-based methods. DL methods published in the last few years can localize 3D cephalometric landmarks with great accuracy, often under the 2-mm threshold of clinical acceptability

(Lee et al. 2019; O'Neil et al. 2019; Torosdagli et al. 2019; Lang et al. 2020; Ma et al. 2020; Yun et al. 2020; Zhang et al. 2020; Bermejo et al. 2021; Chen et al. 2021; Kang et al. 2021; Liu et al. 2021; Chen, Ma, Liu, et al. 2022). The studies showing the best results usually formulate landmark detection as a regression problem, using landmark heatmap regression methods (Zhang et al. 2020; Chen et al. 2021). However, the evaluation of the published models is often limited to few landmarks, and both systematic reviews noted a high risk of bias in the reporting of these studies, mainly because the description of the database/reference was limited and because the accuracy scores were calculated from within-sample validation datasets or very small hold-out test sets (<10 scans). As a result, major concerns remain about the robustness and generalizability of DL methods for 3D cephalometric landmarking, highlighting the need for additional evaluation studies with clinically relevant datasets, clear reference data and broader outcome metrics (Dot et al. 2020; Schwendicke, Chaurasia, et al. 2021).

Recently, the fully convolutional neural network (CNN) SpatialConfiguration-Net (SCN) was proposed as a heatmap regression method integrating a spatial configuration module for landmark localization (Payer et al. 2019). SCN has shown impressive results for the localization of anatomic landmarks on datasets of hand radiographs, lateral cephalograms and spine CT scans, but has yet to be evaluated on craniomaxillofacial CT scans (Payer et al. 2019; Sekuboyina et al. 2021). One difficulty to overcome is data size, as high resolution Head CT scans exceed the memory capacity of a typical graphical processing unit (GPU). There are two solutions to overcome this obstacle: 1) downsampling the scans by decreasing their resolution; 2) implementing the CNNs on small 3D image patches. However, downsampled data necessarily result in less accurate landmark localization, while image patches oftentimes lack volumetric context. But a 2-step, coarse-to-fine approach combining both methods could overcome these limitations (Chen et al. 2021; Sekuboyina et al. 2021).

The main goal of this diagnostic accuracy study was to design and implement a coarse-to-fine DL method based on SCN for automatic landmark localization (the index test), before thoroughly comparing its diagnostic performance with respect to manual landmarking (the reference test) on a hold-out test dataset of craniomaxillofacial CT scans from clinical practice.

7.3. Materials and Methods

This study was approved by an appropriate Institutional Review Board (IRB No. CRM-2001-051) and its reporting followed recently published recommendations on artificial intelligence in dental research (Schwendicke, Singh, et al. 2021).

7.3.1. Dataset

Two hundred presurgical CT scans, randomly selected and anonymized, were obtained from a retrospective sample described in a previous study (Dot et al. 2022), consisting of consecutive patients having undergone orthognathic surgery between January 2017 and December 2019 in a single maxillofacial surgery department. Patients referred to this university hospital located in a cosmopolitan European capital city were ethnically diverse and presented a variety of dentofacial deformities within the scope of orthognathic surgery (maxilla and/or mandible surgery, usually performed along with orthodontic treatment). Two subjects refused to participate; their data was excluded from the dataset. 198 subjects (198 anonymized presurgical CT scans) were eventually included in our dataset and randomly distributed among a training set (*n* = 128), a validation set (*n* = 32) and a test set (*n* = 38) (Supplementary Figure D1). The subjects had a mean age of 27 ± 11 years (minimum age 14, maximum age 60) and 58.6% were females (*n* = 116). 89.4% of the CT scans (*n* = 177) showed metallic artefacts (orthodontic materials, metallic dental fillings or crowns) and 95.4% of the CT scans (*n* = 189) were acquired on the same CT machine. The scans had an average of 744 slices with a mean in-plane pixel size of 0.45*0.45mm^2, mean field of view of 229mm and mean slice thickness of 0.33mm. Full CT scans and patient characteristics are detailed in Supplementary Table D1.

7.3.2. Manual Landmarking (Reference Test)

Thirty-three landmarks, divided into skeletal (*n* = 21) and dental (*n* = 12) landmarks (Fig. 1A), were manually annotated on each CT scan, either once (*n* = 178) by operator #1 (a trained orthodontist with 5 years of clinical experience), or twice (*n* = 20) by operators #1, #2 (a trained orthodontist with 5 years of clinical experience) and #3 (a final year postgraduate maxillofacial surgeon). The reference data used to train and test our DL model were either the single annotations (*n* = 178) or the average of the 6 annotations (*n* = 20, test set only). The scans annotated six times were part of a previous repeatability and reproducibility (R&R) study, which could be used to evaluate intra and interobserver variability of the reference test (Dot et al. 2021). In the test set, some CT scans showed missing dental landmarks: 16O (*n* = 1), 26O (*n* = 1), 31A (*n* = 2), 31E (*n* = 2), 36O (*n* = 1), 46O (*n* = 2). Landmark definitions and landmarking procedure are detailed in the Supplementary Materials D.

7.3.3. Deep Learning-Based Landmarking (Index Test)

The DL model implemented in this study was the publicly available SCN described by Payer et al. (Payer et al. 2019), running in Tensorflow v1.15.0 on our laboratory workstation (CPU AMD Ryzen 9 3900X

12-Core; 128 Gb RAM; GPU Nvidia Titan RTX 24Gb). The pipeline used to train the network followed a coarse-to-fine approach: 1) to keep most of the volumetric context, we trained a first network (SCN#1) on downsampled-resolution full scans; 2) to localize the landmarks more accurately within selected regions of interest (ROIs), five networks (SCN#2 to SCN#6) were trained on selected full-resolution ROIs (Fig. 1B). The coordinates of each local heatmap maxima were considered as the predicted landmark positions. The confidence in a network prediction was evaluated as "very low" when the heatmap maximum value was below a threshold established from the validation results. Please refer to the Supplementary Materials D for additional implementation details.

Inference (prediction made by the trained model) was performed on our hold-out test set (n = 38) following a 2-stage method (Fig. 1B). At stage 1, SCN#1 predicted the "coarse" localization of the landmarks, which was then used to extract the 5 ROIs. At stage 2, SCN#2 to SCN#6 predicted the "fine" localization of the landmarks in each ROI along with the confidence in the prediction. This method systematically localized 33 landmarks for each CT scan. In CT scans with missing landmarks (*i.e.* missing teeth), the corresponding predictions were considered as missing values and deleted by the operator.

7.3.4. Evaluation

If a CT scan showed several "very low" confidence levels in coordinate predictions, the subject was considered as an outlier case. To evaluate the overall localization performance on our test set, three commonly-used criteria were computed for each landmark (Wang et al. 2016): 1) mean radial error (MRE) – mean Euclidian distance between the reference landmark and the predicted landmark; 2) success detection rate (SDR) – proportion of landmarks located with radial errors under 2mm, 2.5mm, 3mm; 3) minimum and maximum radial error.

Conventional 2D cephalometric measurements (Supplementary Table D4) were computed using orthogonal projections of the 3D landmarks on a midsagittal plane computed using a previously published automated method (Pinheiro et al. 2019). Additionally, the accuracy of Frankfort horizontal (FH) plane construction (porion right/left and orbitale left) was evaluated.

The predicted landmarks and cephalometric variables were compared to the Bland-Altman 95% limits of agreement (LoA) of manual landmarking and cephalometric measurement reproducibility, computed from a previous R&R study (Dot et al. 2021) following ISO norm 5725 (ISO 5725-2:2019). More details are provided in the Supplementary Materials D.

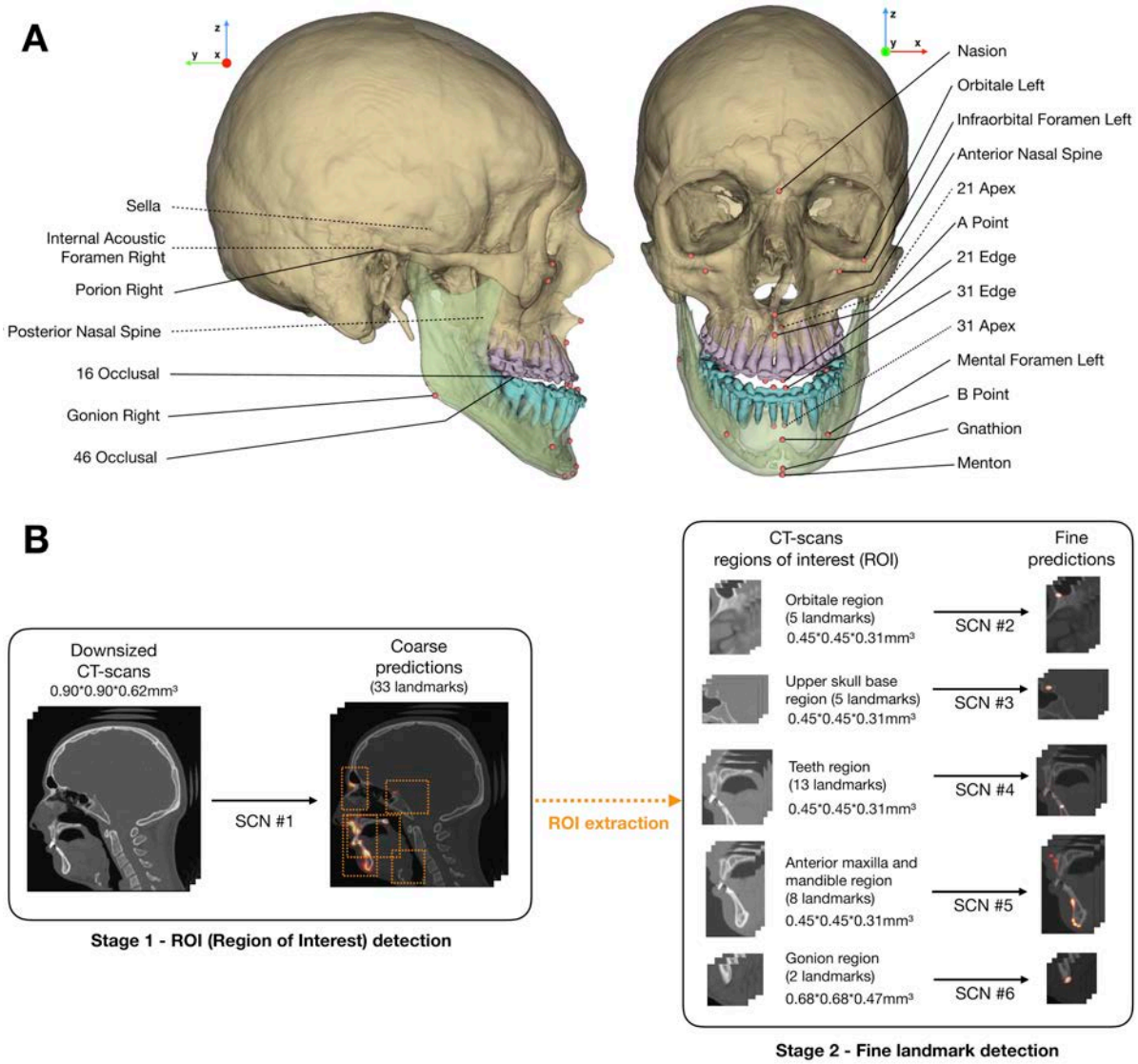


Figure 43: Landmarks and pipeline of the deep learning model. (A) Illustration of the set of 33 landmarks; bilateral landmarks are named once; dotted lines show landmarks localized inside the skull; (B) 2-stage method used for model inference. SCN, SpatialConfiguration-Net; ROI, region of interest.

7.3.5. Statistical analysis

Continuous variables were presented as means ± standard deviations; categorical variables were expressed as numbers and percentages. We first assessed the normality of the data using Shapiro-Wilk normality test, and then applied Wilcoxon and Student t-tests for nonparametric and parametric data, respectively; p-values < 0.05 were considered statistically significant.

7.4. Results

7.4.1. Training, testing and outlier case

Training time for one network on one GPU was about 48 hours and inference required around 1 minute per CT scan. One CT scan from a patient exhibiting cleidocranial dysplasia showed several predictions (A Point and several dental landmarks) with “very low” confidence levels. It was therefore considered as an outlier case and was excluded from the overall analysis, although individual localization performance was assessed and discussed.

7.4.2. Localization performance

On our test set without the outlier case (*n* = 37), MRE for all landmarks was 1.0mm ± 1.3mm and SDRs for all landmarks were 90.4%, 93.6% and 95.4%, using 2mm, 2.5mm and 3mm precision ranges, respectively (Table 1). Thirteen landmarks (39.4%) showed SDRs at 2mm of 100%; 24 landmarks (72.7%) showed SDRs at 2mm over 90%, and 5 landmarks (15.2%) showed SDRs at 2mm under 80% (B point, gonion left and right, orbitale left and right). Additional results, including the outlier case and validation set evaluations, are reported in Supplementary Tables D6, D7 and D8. When comparing scans with references constructed from 1 or 6 annotations, 3 landmarks exhibited statistically significantly larger errors when constructed from 6 annotations instead of 1 annotation: orbitale left, 11 incisal edge and 41 incisal edge.

7.4.3. Cephalometric Analysis

On our test set without the outlier case (*n* = 37), mean differences between the reference and predicted measurements were -0.3 ± 1.3° for angular observations and -0.1 ± 0.7mm for linear observations (Table 2). 96.7% (*n* = 322) of the skeletal measurements and 83.8% (*n* = 181) of the dentoalveolar measurements showed errors inferior to 2mm/2°. The mean absolute angular distance between predicted and reference FH planes was 0.4 ± 0.3°, and all the measurements were inferior to 2°.

Table 19: Mean radial errors (mm), success detection rates (% (n)) and minimum/maximum radial error (mm) for each landmark on the hold-out test set without the outlier case (n = 37). MRE, mean radial error; SD, standard deviation; Min., minimum radial error; Max., maximum radial error; L, left; R, right.

	MRE ± SD	<2mm	<2.5mm	<3mm	Min.	Max.
11 Apex	0.7 ± 0.4	100 (37)	100 (37)	100 (37)	0.2	1.5
11 Edge	0.4 ± 0.3	100 (37)	100 (37)	100 (37)	0.1	1.3
16 Occlusal	1.3 ± 2.4	94.4 (34)	94.4 (34)	94.4 (34)	0.1	11.2
21 Apex	0.7 ± 0.3	100 (37)	100 (37)	100 (37)	0.2	1.9
21 Edge	0.5 ± 0.3	100 (37)	100 (37)	100 (37)	0.1	1.4
26 Occlusal	1.2 ± 2.4	94.4 (34)	94.4 (34)	94.4 (34)	0.1	11.4
31 Apex	0.9 ± 1.4	97.1 (34)	97.1 (34)	97.1 (34)	0.2	8.7
31 Edge	0.6 ± 1.1	94.4 (33)	97.1 (34)	97.1 (34)	0.1	6.7
36 Occlusal	1.5 ± 2.9	91.7 (33)	91.7 (33)	91.7 (33)	0.2	11.3
41 Apex	0.6 ± 0.3	100 (37)	100 (37)	100 (37)	0.2	1.3
41 Edge	0.5 ± 0.2	100 (37)	100 (37)	100 (37)	0.1	1.3
46 Occlusal	0.9 ± 1.8	97.2 (35)	97.2 (35)	97.2 (35)	0.6	11.0
A Point	1.1 ± 0.9	89.2 (33)	91.9 (34)	91.9 (34)	0.2	3.9
Anterior Nasal Spine	0.7 ± 0.7	94.6 (35)	94.6 (35)	97.3 (36)	0.1	3.2
B Point	1.7 ± 1.5	67.6 (25)	81.1 (30)	91.9 (34)	0.3	8.5
Gnathion	1.6 ± 0.6	91.9 (34)	97.3 (36)	100 (37)	0.3	2.5
Gonion L	1.9 ± 1.7	70.3 (26)	75.7 (28)	86.5 (32)	0.3	7.3
Gonion R	2.1 ± 1.4	48.7 (18)	70.3 (26)	73.0 (27)	0.3	6.9
Infraorbital Foramen L	0.6 ± 0.3	100 (37)	100 (37)	100 (37)	0.2	2.0
Infraorbital Foramen R	0.6 ± 0.5	97.3 (36)	100 (37)	100 (37)	0.1	2.4
Internal Acoustic Foramen L	0.6 ± 0.4	100 (37)	100 (37)	100 (37)	0.2	1.9
Internal Acoustic Foramen R	0.6 ± 0.6	97.3 (36)	97.3 (36)	97.3 (36)	0.1	3.9
Mental Foramen L	0.4 ± 0.2	100 (37)	100 (37)	100 (37)	0.1	0.8
Mental Foramen R	0.4 ± 0.3	100 (37)	100 (37)	100 (37)	0.1	1.3
Menton	1.6 ± 0.6	94.6 (35)	97.3 (36)	100 (37)	0.4	2.6
Nasion	0.6 ± 0.3	100 (37)	100 (37)	100 (37)	0.1	1.9
Orbitale L	2.7 ± 2.0	43.2 (16)	56.8 (21)	67.6 (25)	0.1	8.8
Orbitale R	2.6 ± 2.3	56.8 (21)	67.6 (25)	70.3 (26)	0.3	9.7
Pogonion	1.1 ± 0.6	89.2 (33)	97.3 (36)	100 (37)	0.2	3.0
Porion L	1.1 ± 0.5	89.2 (33)	100 (37)	100 (37)	0.2	2.3
Porion R	1.3 ± 0.7	86.5 (32)	89.2 (33)	100 (37)	0.3	2.8
Posterior Nasal Spine	0.5 ± 0.4	100 (37)	100 (37)	100 (37)	0.1	1.5
Sella	0.8 ± 0.4	100 (37)	100 (37)	100 (37)	0.2	2.0

Table 20: Mean errors (mm) and success detection rates (% (n)) for each cephalometric variable on the hold-out test set without the outlier case (n = 37). SD, standard deviation.

	Mean	SD	<2mm/2°
Skeletal			
SNA (°)	-0.1	0.7	100 (37)
SNB (°)	-0.0	0.7	100 (37)
ANB (°)	-0.1	0.2	100 (37)
ANS-PNS / Go-Gn (°)	-0.1	1.3	91.9 (34)
S-Na / Go-Gn (°)	0.0	1.4	83.8 (31)
Pog to NB (mm)	0.0	0.4	100 (37)
A to MSP (mm)	0.0	0.3	100 (37)
B to MSP (mm)	-0.3	0.6	97.3 (36)
Pog to MSP (mm)	-0.2	0.7	97.3 (36)
Dentoalveolar			
SN / Occlusal plane (°)	-0.2	1.2	97.1 (34)
Upper inc / ANS-PNS (°)	-0.8	1.5	75.7 (28)
Upper inc to NA (mm)	0.2	0.5	100 (37)
Inter-incisal angle (°)	-1.1	1.9	54.3 (19)
Lower inc / Go-Gn (°)	-0.4	1.8	77.1 (27)
Lower inc to NB (mm)	-0.1	1.0	97.3 (36)

7.4.4. Comparison with manual landmarking and measurement reproducibility

On our test set without the outlier case (*n* = 37), when comparing predicted landmark coordinates in the *-x*, *-y* and *-z* directions with manual landmarking repeatability, 90.7% (*n* = 2114*)* of the skeletal coordinates and 65.4% (*n* = 871) of the dental coordinates were within 95% LoA (Supplementary Table D9). When comparing predicted cephalometric measurement errors with manual measurement repeatability, 91.9% (*n* = 306) of the skeletal variables and 71.8% (*n* = 155) of the dentoalveolar variables were within 95% LoA (Supplementary Table D10). For the scans included in the R&R study (without the outlier case, *n* = 19), localization and measurement error boxplots for the manual and automatic methods are shown in Figure 2. Bland-Altman plots showing the deviations of manual and automatic landmarking localizations and cephalometric measurements for the scans included in the R&R study are reported in the Supplementary Materials D. Automatic localization errors were statistically significantly larger than manual landmarking errors (Fig. 2) for 5 skeletal landmarks (32.8%), 10 dental landmarks (83.3%), 1 skeletal measurement (11.1%) and 2 dentoalveolar measurements (33.3%).

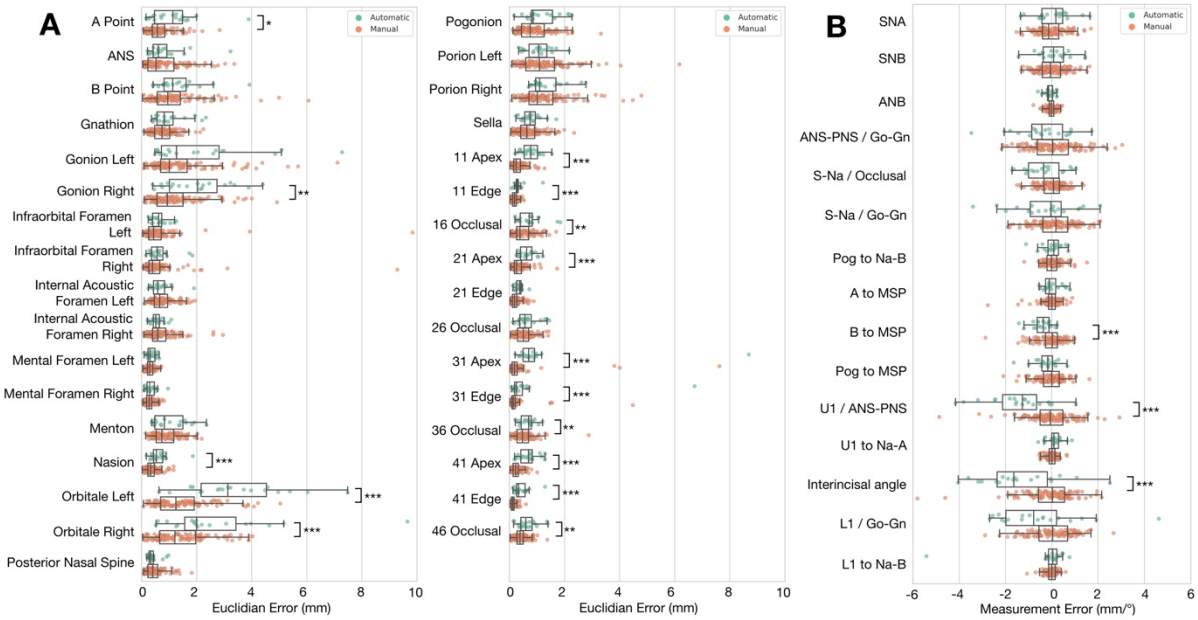


Figure 44: Localization and measurement error boxplots for automatic (green) and manual (orange) methods on 19 CT scans from the test set. (A) Localization errors (mm) for each landmark; (B) Measurement errors (mm/°) for each cephalometric variable. For each pair of results, statistically significant differences are indicated (* p<0.05; ** p<0.01; ***p<0.001).

7.4.5. Three-Dimensional Visualization

We chose two subjects representative of our test dataset as well as the “outlier case” to illustrate our results. Figure 3 shows reference and predicted landmarks plotted on the fully automatically-obtained CT scan segmentations (Dot et al. 2022).

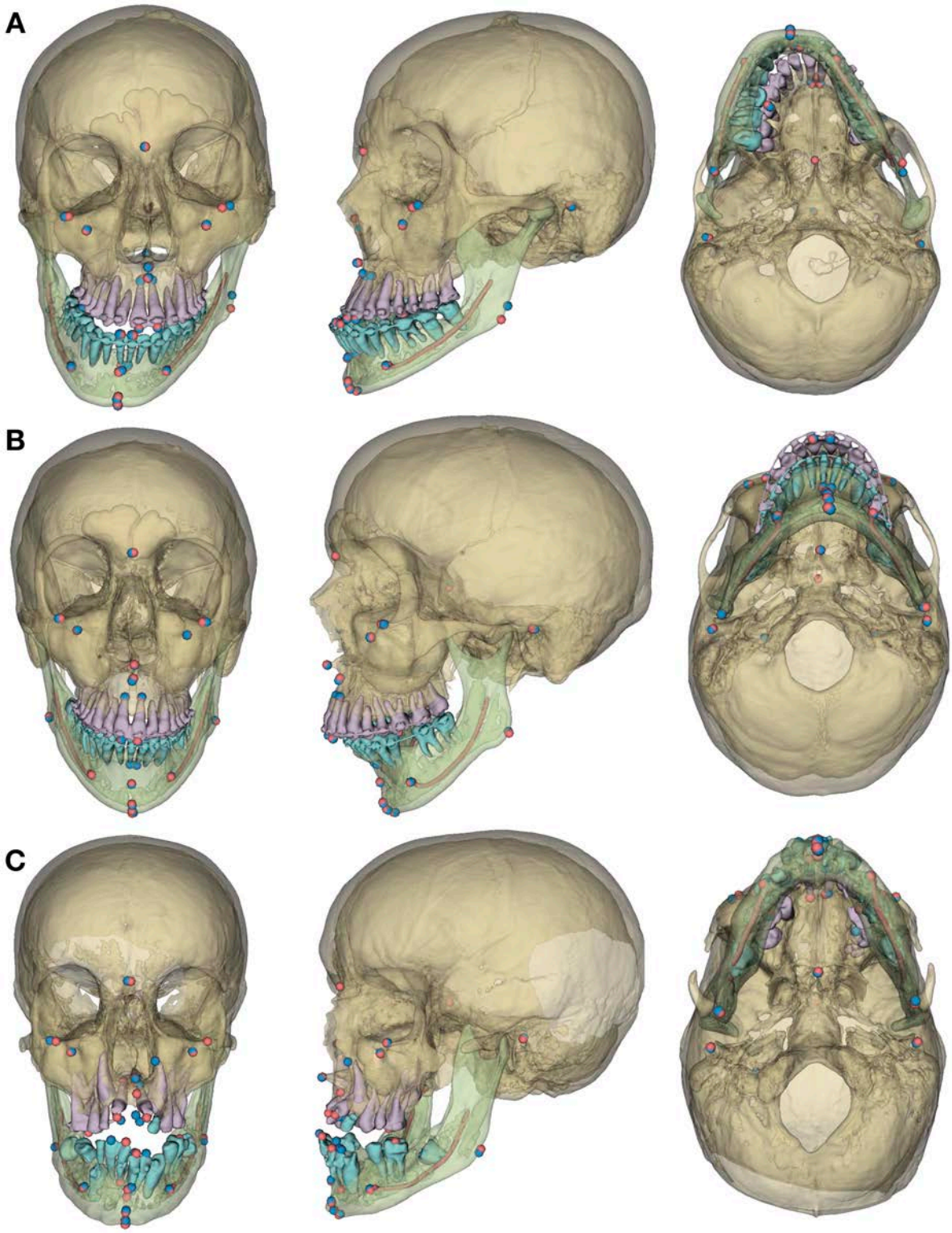


Figure 45: Frontal, ¾ left and inferior views of the 3D models, reference (red) and predicted (blue) landmarks for 3 subjects. (A) Prognathic and asymmetric mandible; (B) retrognathic mandible; (C) craniofacial syndrome “outlier case”, the errors in the predicted A point (at the level of the upper left canine apex) and the dental landmarks are to be noted.

7.5. Discussion

The increasingly common use of 3D scans to assess complex maxillomandibular deformities and to plan orthognathic surgeries implies a critical need for clinical implementation of 3D cephalometric analyses. Such analyses currently require manual localization of 3D landmarks, a task that is time-consuming (± 15mn) and demands highly trained operators. In this study, we trained a DL network in order to localize 33 cephalometric landmarks automatically before evaluating the model on a challenging hold-out test set from clinical practice. The proposed DL pipeline took around one minute to localize the landmarks in a fully automatic manner. This amounts to a significant reduction of the time and effort needed for the task. The landmarks were localized with high accuracy, with 90.4% less than 2mm away from the manually localized reference landmarks.

Heterogeneity in the methods and datasets make studies reporting DL results notoriously difficult to evaluate and compare (Schwendicke, Singh, et al. 2021). The main strength of our study is that it provides a validation of our method based on a clinically relevant test dataset, randomly selected from a clinical sample of presurgical CT scans, 92.1% of which had metal artifacts. Moreover, we carefully constructed our reference test (manual landmarking) using the means of the six repetitions from a previously published R&R study for twenty of the test scans and asking one of this R&R study's operators to label the 178 remaining scans. Overall, our results are comparable to current state-of-the art studies localizing landmarks on CBCT scans, some landmarks showing slightly better and other slightly worse localization results (Torosdagli et al. 2019; Zhang et al. 2020; Chen et al. 2021). However, previous studies lacked a clear definition of their dataset, localized fewer landmarks and evaluated their results following a cross-validation approach with no hold-out test dataset, which might question the generalizability of the results. It must be noted that our study focused on CT scans because it is the only imaging modality used for computer-assisted planning and personalized implant manufacturing for orthognathic surgery in our maxillofacial surgery department at this time. In future works we plan to use CBCT data in order to fine-tune our model and evaluate its accuracy on this other widespread imaging modality. Currently, our method does not perform automatic detection of the presence or absence of the landmarks; in the case of missing landmarks, those were deleted manually. We considered this approach sufficient, as it is easy for an operator to identify missing landmarks when running the cephalometric analysis, but other methods have been suggested to perform this task automatically (Lang et al. 2020; Chen et al. 2021).

The main goal of cephalometric landmarking is to perform linear and angular measurements which will ultimately provide clinical guidance. In order to evaluate the clinical usefulness of our DL-based

method, our outcome set included lateral cephalometric measurements commonly found in R&R studies (van Bunningen et al. 2022) as well as three additional frontal measurements. We chose to perform 2D cephalometric measurements based on orthogonally projected landmarks because 3D cephalometric analysis remains complex, and thus beyond the scope of this study (Gateno et al. 2011). These measurements do not use the full potential of 3D cephalometry, but we believe they provide useful insight on the potential clinical usefulness of the method.

Concerning the skeletal landmarks, it has been shown that the reproducibility of manual landmarking was highly dependent on the type of landmark: landmarks localized on clear anatomical boundaries (*e.g.*, sutures, spikes, holes) tend to be more reproducible than landmarks localized on skeletal contours (Sam et al. 2018). The comparison of our DL-based method with manual landmarking reproducibility shows that it is on par with trained clinicians for the localization of skeletal landmarks. Error-prone landmarks tend to be the same whether the landmarking is performed manually or automatically. Furthermore, even the landmarks with the worst accuracy results provided highly accurate cephalometric measurements or FH plane constructions, comparable with those obtained by clinicians. This confirms the need for evaluation outcomes other than MRE and SDR, as radial errors do not necessarily translate into clinically relevant errors (Gupta et al. 2016). Interestingly, landmarks localized on the craniofacial foramens showed excellent accuracy results, with 99.1% (n = 220) of the landmarks located within 2mm from the reference. These "novel" landmarks, which could not be localized on 2D cephalograms, could be used in future 3D cephalometric analyses (Naji et al. 2014; Lim et al. 2019; Dot et al. 2021).

Concerning the dental landmarks, despite good overall accuracy, the automated method provided less reliable results than the clinicians, with several automatic localizations showing errors statistically significantly larger than manual localization errors. The localization of these landmarks could probably be improved by refining their positions on the CT scan segmentation, for example using an additional knowledge-based method (Montúfar et al. 2018). When the patients' intraoral scans are superimposed on the CT scans, for surgery planning for instance, they may also be segmented automatically and used for refining crown landmark localization (Hao et al. 2021 Nov 1).

We excluded one subject showing several landmarks with "very low" confidence levels, because such levels usually signal that the network did not work as expected and could lead to major errors. In this case, several landmarks (A Point and dental landmarks) showed errors >10mm (Supplementary Table D7) and required operator corrections. These errors are probably due to the atypical anatomy of this subject, who exhibited a rare syndromic disease with several included teeth (Fig. 3C). From a

clinical viewpoint, additional verification and correction of the results could be performed on a visualization of the predicted landmarks plotted on 3D models obtained fully automatically via DL (Fig. 3) (Wang et al. 2021; Dot et al. 2022).

To conclude, the proposed method achieved high accuracy on a test set of presurgical CT scans, providing results on par with those of clinicians for skeletal landmark localization and subsequent cephalometric measurements. The localization of dental landmarks still requires improvement to provide more reliable cephalometric measurements. Despite these promising results, our model requires additional testing in order to further evaluate its generalizability, reproducibility and robustness outside the scope of the present dataset. The data augmentation procedure that we applied during model training, based on image manipulations, should be helpful for the generalizability of the model (Shorten and Khoshgoftaar 2019) but it still has to be evaluated on an external test dataset including data from other clinical centers and CT machines. Afterwards, a prospective diagnostic efficacy study should evaluate the impact of using such an automated tool in routine clinical practice.

PARTIE D :

Perspectives

8 : Illustration de l'intérêt clinique des résultats pour la planification de chirurgie orthognathique

8.1. Problématique

Dans ce dernier chapitre, nous souhaitons illustrer l'intérêt et l'applicabilité clinique de nos résultats dans le contexte de la planification de la chirurgie orthognathique (Figure 15). Lorsque le patient est prêt pour la chirurgie (le plus souvent après une phase de traitement orthodontique), les cliniciens prenant en charge le patient (chirurgien, et orthodontiste le cas échéant) le reçoivent pour effectuer un bilan pré-chirurgical comprenant son anamnèse, son examen clinique (visage et tissus mous, occlusion, etc.) et l'étude de ses moulages dentaires et examens radiographiques. A partir de ces éléments, les cliniciens planifient le type de chirurgie (du maxillaire et/ou de la mandibule et/ou du menton habituellement), la direction et la quantité des déplacements à effectuer. La planification chirurgicale en 3D par ordinateur est réalisée à partir d'une imagerie 3D (CT-Scan ou CBCT) et des moulages dentaires numérisés, après segmentation et annotation de points céphalométriques sur ces données. Pour transposer cette planification au bloc opératoire, des dispositifs chirurgicaux sur-mesure (gouttières occlusales, guides chirurgicaux, plaques d'ostéosynthèse, etc.) sont produits à partir de ces éléments numériques.

La planification chirurgicale par ordinateur peut être effectuée directement au sein de l'établissement de soin, par exemple en utilisant des logiciels commerciaux et en imprimant des gouttières occlusales. Dans ce cadre, nos travaux sur l'automatisation du traitement des imageries CT-Scan pourraient permettre d'alléger le procédé de planification numérique 3D, en automatisant la segmentation et le placement des points céphalométriques. Cependant, les contraintes réglementaires et techniques pour la réalisation de ce type de planification en interne sont importantes, ce qui explique que de nombreux cliniciens travaillent en collaboration avec un partenaire industriel.

Dans le cas d'une chirurgie planifiée avec un partenaire industriel, les cliniciens transmettent une prescription des déplacements à effectuer au partenaire industriel, qui va effectuer une proposition de planification chirurgicale correspondante à partir des imageries et des moulages dentaires. Cette proposition de planification est ensuite soumise aux cliniciens, qui vont l'amender. Après validation de la planification finale par les cliniciens, la production des dispositifs chirurgicaux sur-mesure est initiée par le partenaire industriel (Figure 46).

Actuellement, la segmentation et le placement des points céphalométriques sur les imageries 3D sont donc souvent externalisés auprès d'un expert industriel. Les cliniciens obtiendront ces éléments plusieurs jours après la consultation pré-chirurgicale, les encourageant à utiliser des méthodes classiques de céphalométrie 2D pour déterminer le plan de traitement chirurgical pendant qu'ils sont en présence du patient. Une fois en possession des éléments 3D, les cliniciens sont le plus souvent amenés à amender la planification proposée par le partenaire industriel. Ainsi, la prescription initiale des déplacements à effectuer serait très probablement améliorée si les cliniciens pouvaient visualiser lors de leur bilan pré-chirurgical les modèles 3D du patient et directement effectuer une analyse céphalométrique 3D. Nos travaux sur l'automatisation du traitement des imageries CT-Scan pourraient ainsi permettre d'alléger le procédé de planification numérique 3D réalisé avec un partenaire industriel (Figure 46).

Nous présenterons 3 cas cliniques pour illustrer comment les modèles d'apprentissage profond présentés dans ce manuscrit pourraient être utilisés dans ce contexte. Ces 3 cas cliniques ont été sélectionnés parmi les patients ayant récemment reçu une planification chirurgicale dans le service de chirurgie maxillo-faciale de la Pitié-Salpêtrière, et ne sont donc pas inclus dans les bases de données précédemment présentées.

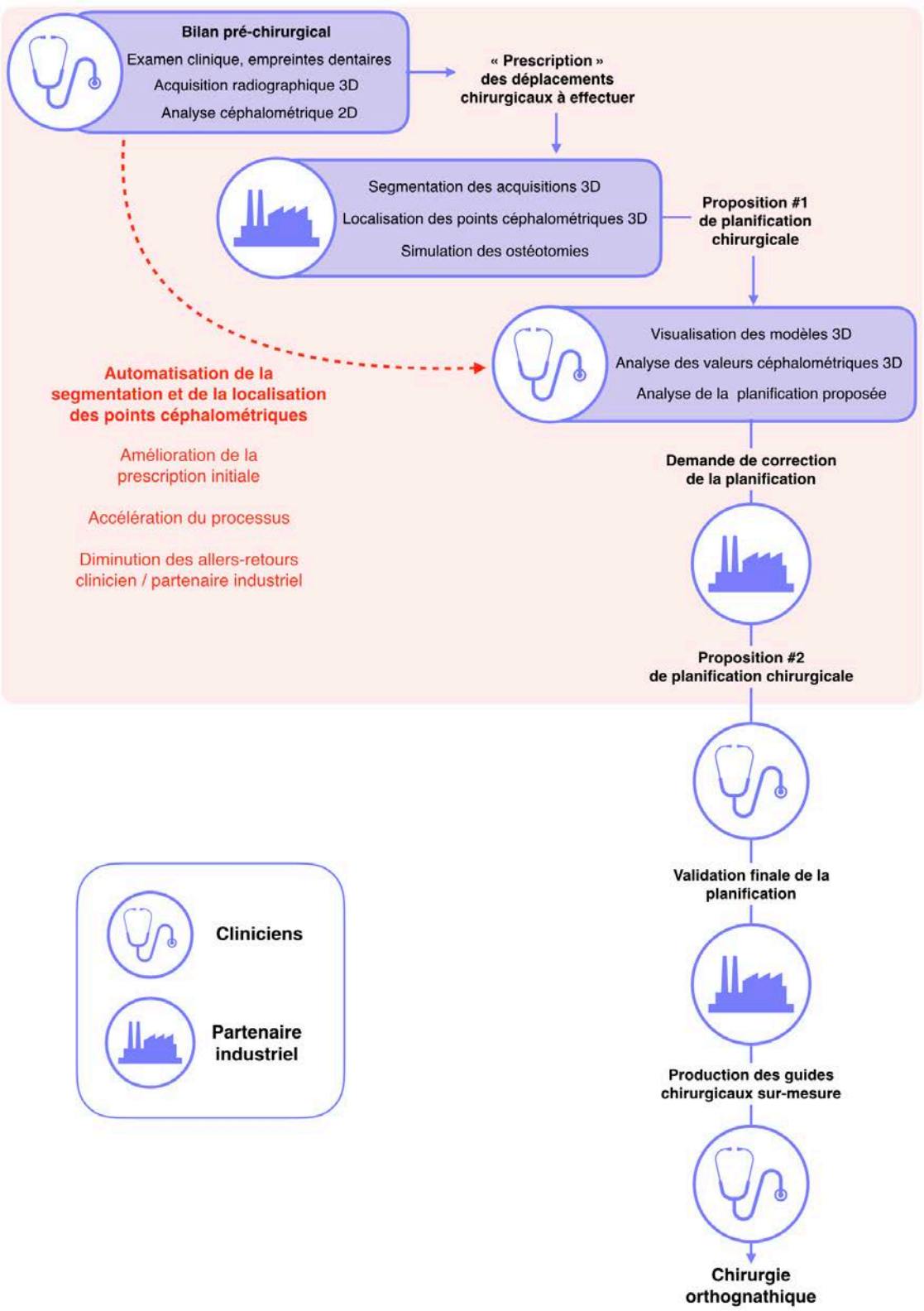


Figure 46 : Représentation schématique d'un processus de planification chirurgicale 3D numérique pour la conception de guides chirurgicaux sur-mesure avec un partenaire industriel, illustrant les relations entre les cliniciens et le partenaire industriel.

8.2. Cas cliniques

8.2.1. Cas clinique 1

Le premier cas clinique est celui d'un patient de 20 ans présentant un décalage squelettique de classe III, une béance antérieure et une asymétrie mandibulaire (Figure 47).

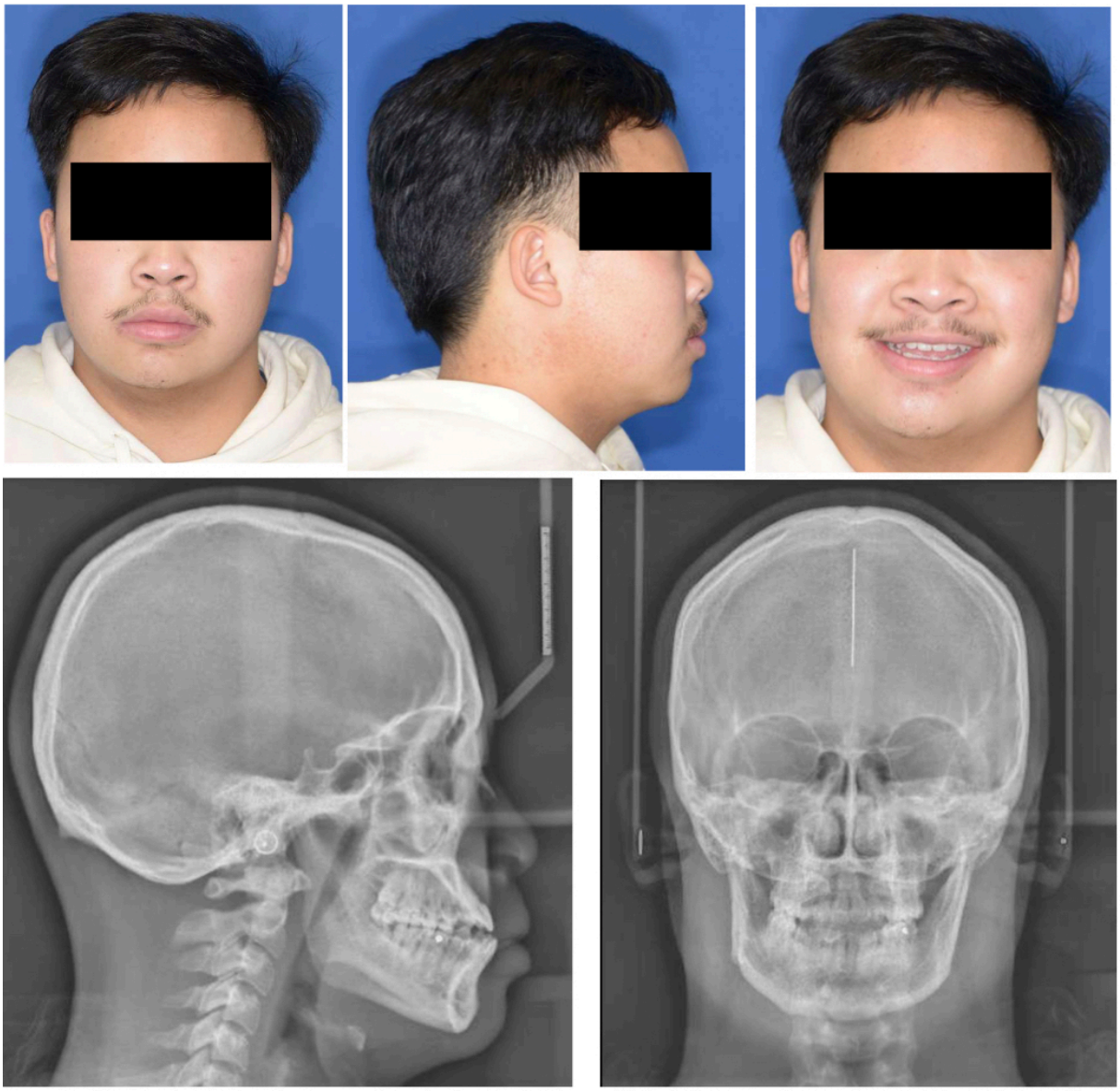


Figure 47 : Cas clinique 1, photographies exo-buccales et examens radiographiques 2D.

Nos modèles d'apprentissage profond (Dot et al. 2022; Dot et al. 2022 Aug 18) ont permis d'effectuer de façon entièrement automatisée la segmentation et le placement des points céphalométriques 3D sur les imageries pré-chirurgicales CT-Scan (Figure 48).

Figure 48 : Cas clinique 1, segmentation automatisée des imageries pré-chirurgicales CT-Scan et placement automatisé des points céphalométriques 3D. Une erreur au niveau de la localisation du point 21E (superposé sur le point 11E) est à noter.

A partir de ces éléments, une analyse géométrique peut être effectuée afin de localiser plus finement le siège de l'asymétrie. Nous proposons de relier certains des points céphalométriques, afin de visualiser les anomalies anatomiques (Figure 49) :

- les axes médians de la base crânienne et du maxillaire peuvent être visualisés en reliant les points Sella (S), Nasion (Na), Anterior Nasal Spine (ANS), A et Posterior Nasal Spine (PNS) ;
- le plan de Francfort peut être visualisé en reliant les points Porion (Po) droit, Po gauche, Orbitale (Or) droit et Or gauche ;
- l'axe médian de la mandibule et le plan sous mandibulaire peuvent être visualisés en reliant les points B, Pogonion (Pog), Gnathion (Gn), Menton (Me), Gonion (Go) droit et Go gauche.

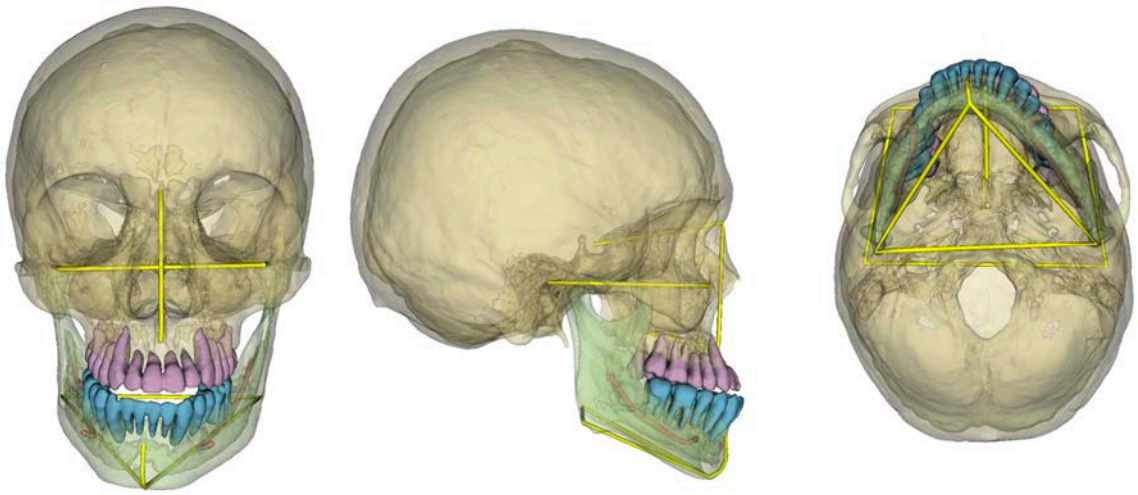


Figure 49 : Cas clinique 1, analyse géométrique globale. L'asymétrie de la mandibule est à noter.

Une analyse géométrique du massif crânien supérieur et des dents maxillaires permet de visualiser la position du plan de Francfort par rapport au plan occlusal maxillaire (en reliant les points 11E, 16O et 26O) et à l'axe incisif (en reliant les points 11E et 11A), comme illustré en Figure 50. On observe ici une légère bascule du plan occlusal vers le haut à droite.

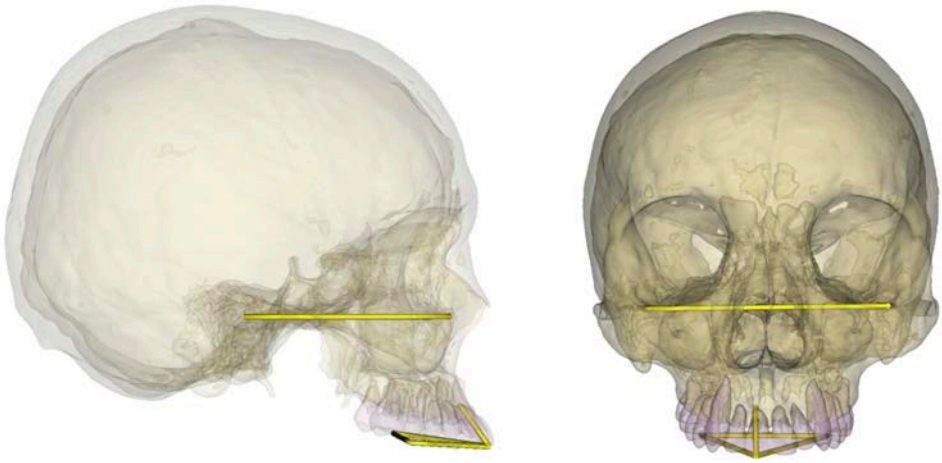


Figure 50 : Cas clinique 1, analyse géométrique analyse géométrique du massif crânien supérieur et des dents maxillaires.

Une analyse géométrique centrée sur la mandibule permet de vérifier que les dents sont globalement bien situées dans le corps mandibulaire en vue occlusale (Figure 51). Pour faciliter la visualisation, nous pouvons relier les points dentaires 41E, 31E, 36O et 46O.

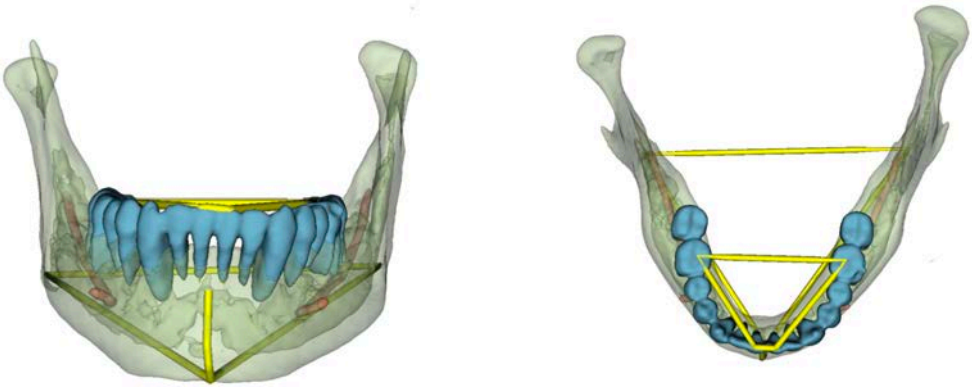


Figure 51 : Cas clinique 1, analyse géométrique de la mandibule et des dents mandibulaires.

Enfin, des mesures céphalométriques peuvent être effectuées. Ici, l'étude du tangage (correspondant à l'évaluation de profil) et/ou du roulis (correspondant à l'évaluation de la bascule frontale) des angles FMA, SNA et SNB (voir paragraphe 1.1.2) est intéressante pour connaître la nouvelle position souhaitable des bases osseuses (Figure 52). Ces mesures sont résumées dans le Tableau 21, et ont été effectuées en utilisant les extensions « Angle Planes » et « Q3DC » du logiciel libre 3D Slicer (Yatabe et al. 2019).

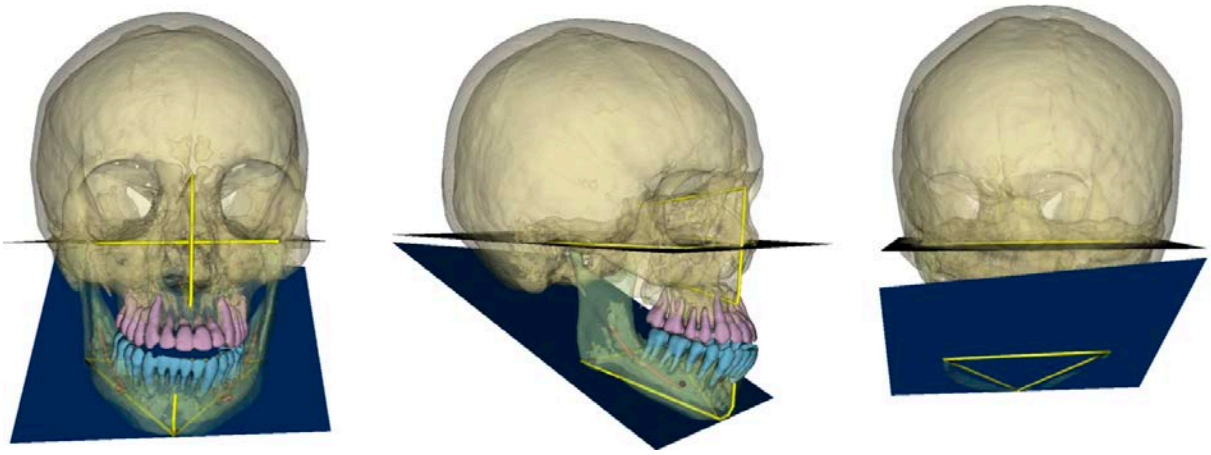


Figure 52 : Cas clinique 1, tracé des plans de Francfort et sous mandibulaire permettant la mesure de l'angle FMA.

Tableau 21 : Cas clinique 1, sélection de mesures céphalométriques 3D

Angle	Plan	Mesure	Éléments diagnostics
FMA	Tangage	25.7°	Typologie normodivergente
	Roulis	4.8°	Augmenté
SNA	Tangage	77.6°	Diminué
SNB	Tangage	81.5°	
ANB	Tangage	-3.9°	Diminué
	Roulis	3.8°	Augmenté
B to S-Na-ANS	Frontal	9.8mm	Menton dévié à droite

Au total, ces données vont dans le sens d'un patient présentant une classe III squelettique par rétromaxillie sur une typologie squelettique normodivergente, avec un plan occlusal basculé en haut à droite et une rotation mandibulaire vers la droite. Ces éléments d'analyse doivent évidemment être confrontés à l'examen clinique pour poser un diagnostic complet, et proposer une planification chirurgicale. La planification chirurgicale effectuée pour ce patient recoupe ces éléments : abaissement à droite et avancée du maxillaire, dérotation de la mandibule (avancée à droite et recul à gauche) (Figure 53).

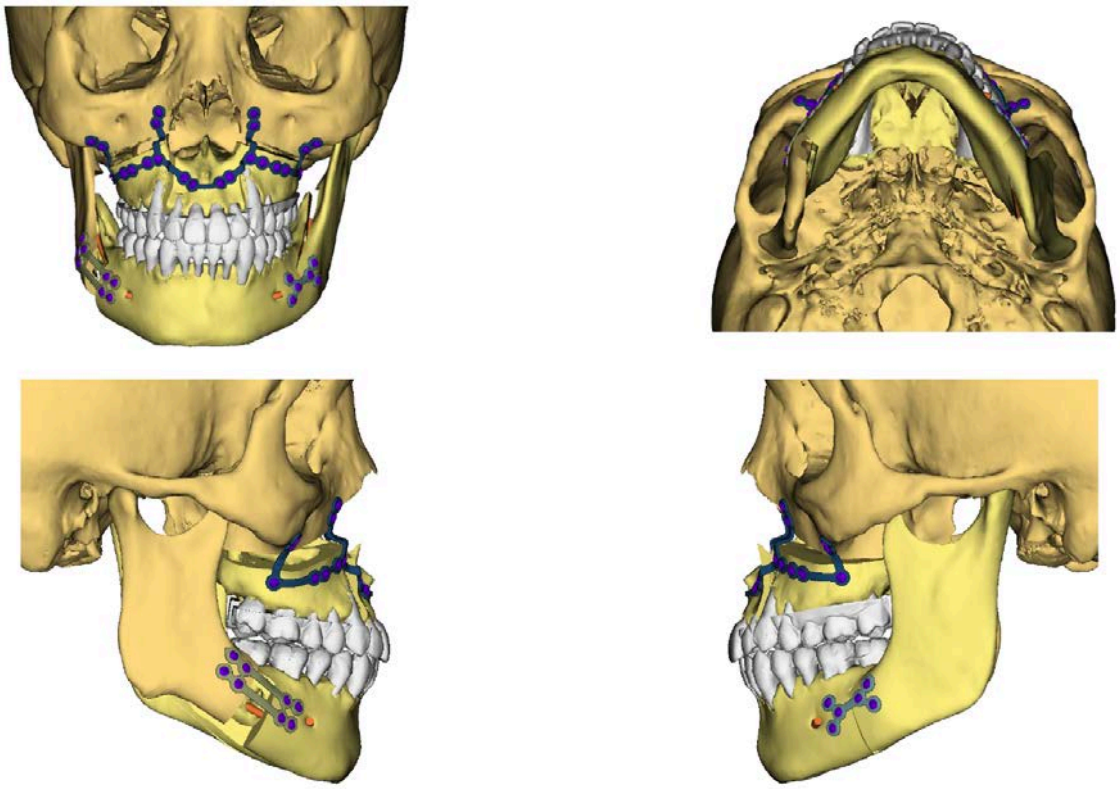


Figure 53 : Cas clinique 1, planification chirurgicale assistée par ordinateur (plaques d'ostéosynthèse sur-mesure).

8.2.2. Cas clinique 2

Le deuxième cas clinique est celui d'un patient de 22 ans présentant un fort décalage squelettique de classe III (Figure 54).

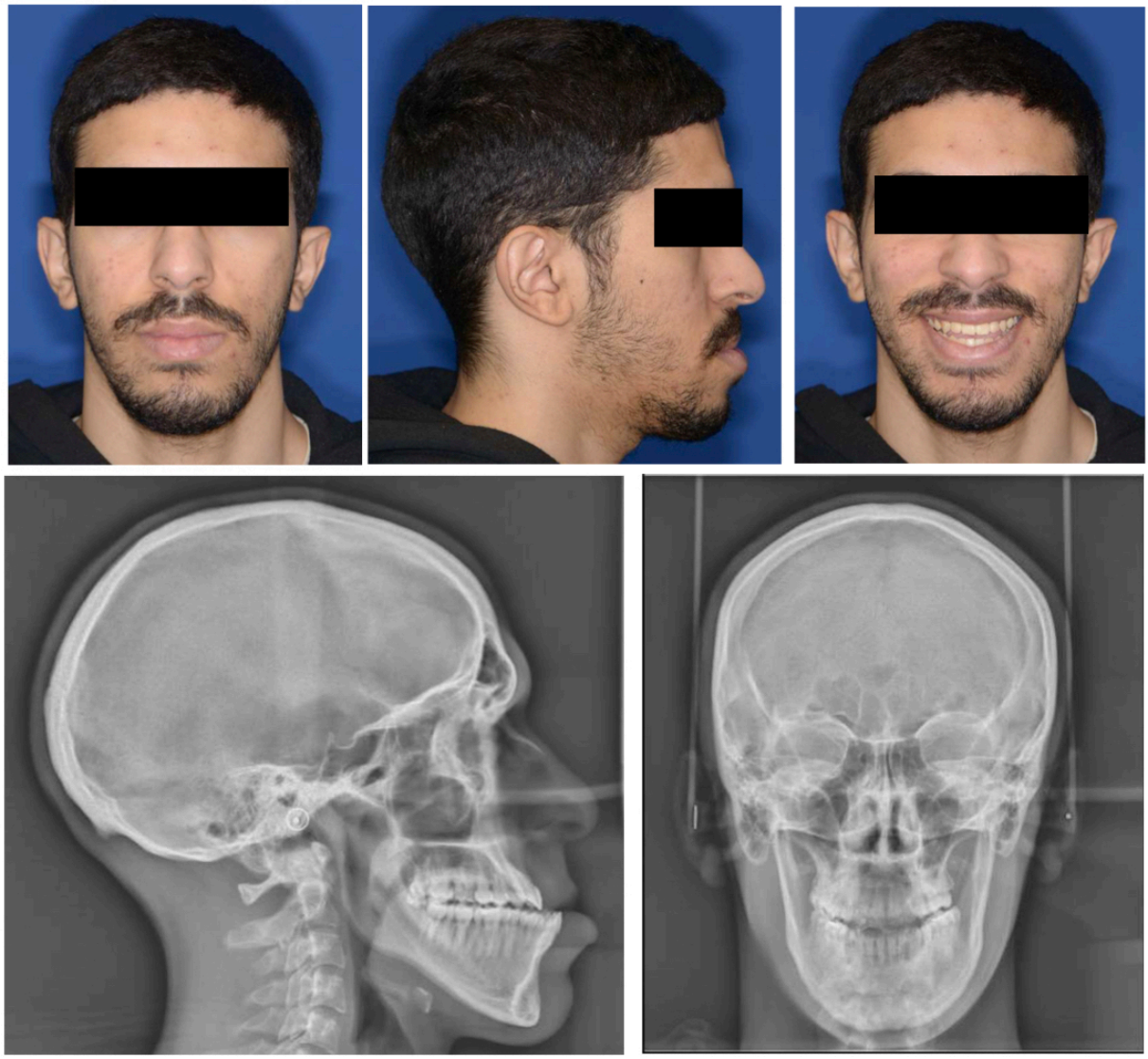


Figure 54 : Cas clinique 2, photographies exo-buccales et examens radiographiques 2D.

Nos modèles d'apprentissage profond ont permis d'effectuer de façon entièrement automatisée la segmentation et le placement des points céphalométriques 3D sur les imageries pré-chirurgicales CT-Scan (Figure 55).

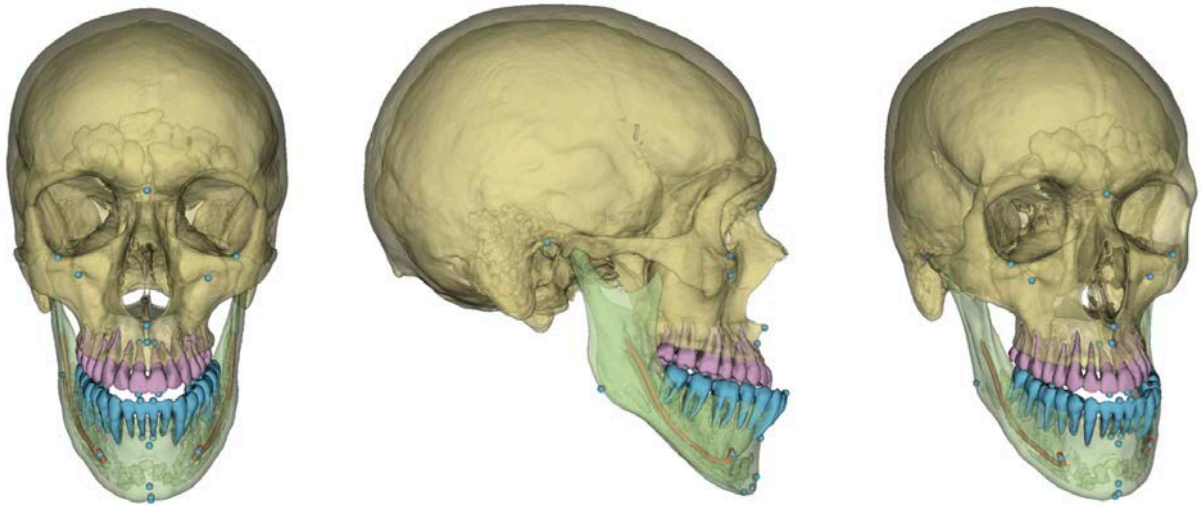


Figure 55 : Cas clinique 2, segmentation automatisée des imageries pré-chirurgicales CT-Scan et placement automatisé des points céphalométriques 3D.

L'analyse géométrique montre que le sujet présente une légère asymétrie mandibulaire (Figure 56). Au niveau des composantes angulaires en vue latérale (tangage), l'angle FMA est augmenté, l'angle ANB est diminué et l'angle SNB est augmenté (Tableau 22).

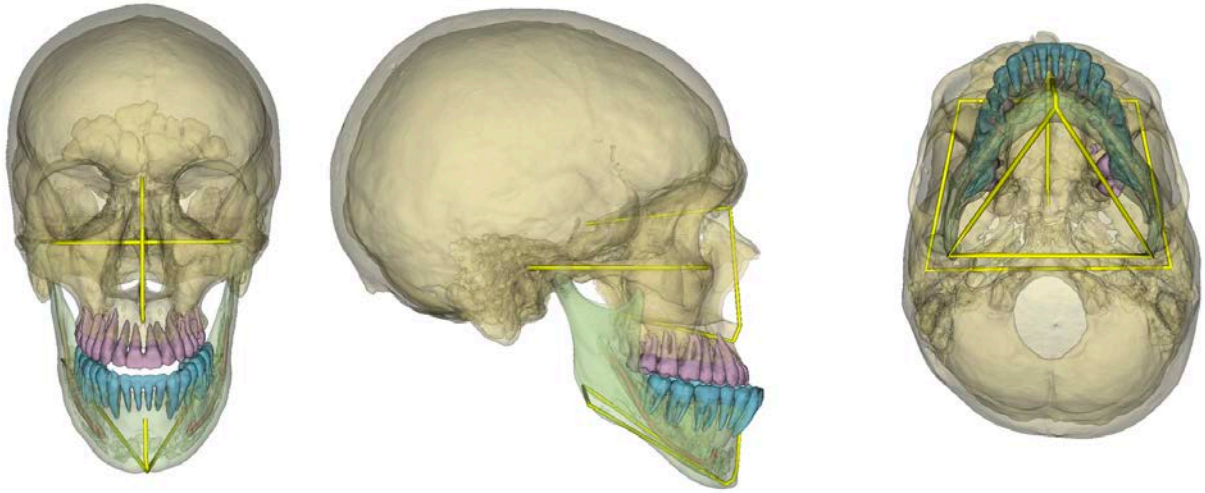


Figure 56 : Cas clinique 2, analyse géométrique globale.

Tableau 22 : Cas clinique 2, sélection de mesures céphalométriques 3D

Angle	Plan	Mesure	Comparaison à la norme (Tweed)
FMA	Tangage	34.5°	Typologie hyperdivergente
	Roulis	0.3°	
SNA	Tangage	81.5°	
SNB	Tangage	82.9°	Augmenté
ANB	Tangage	-1.4°	Diminué

Au total, cette analyse va dans le sens d'un patient présentant une classe III squelettique d'origine mandibulaire sur une typologie squelettique très hyperdivergente, avec une légère déviation de la mandibule vers la gauche. Ces éléments d'analyse doivent évidemment être confrontés à l'examen clinique pour poser un diagnostic complet et proposer une planification chirurgicale. La planification chirurgicale finalement effectuée a consisté en une avancée et impaction maxillaire, un léger recul mandibulaire asymétrique et une génioplastie permettant de diminuer la hauteur verticale antérieure.

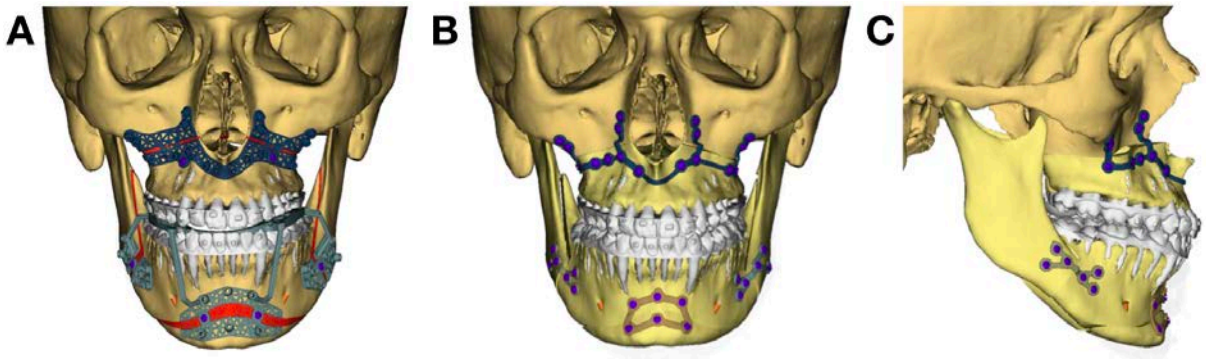


Figure 57 : Cas clinique 2, planification chirurgicale assistée par ordinateur. A. Guides chirurgicaux sur-mesure et traits d'ostéotomies planifiés (en rouge) ; B et C. Plaques d'ostéosynthèse sur-mesure.

8.2.3. Cas clinique 3

Le troisième cas clinique est celui d'une patiente de 40 ans présentant une asymétrie importante (Figure 58). Dans une telle situation, les radiographies 2D de profil et de face sont particulièrement difficiles à interpréter.

Figure 58 : Cas clinique 3, photographies exo-buccales et examens radiographiques 2D.

Nos modèles d'apprentissage profond ont permis d'effectuer de façon entièrement automatisée la segmentation et le placement des points céphalométriques 3D sur les imageries pré-chirurgicales CT-Scan (Figure 59).

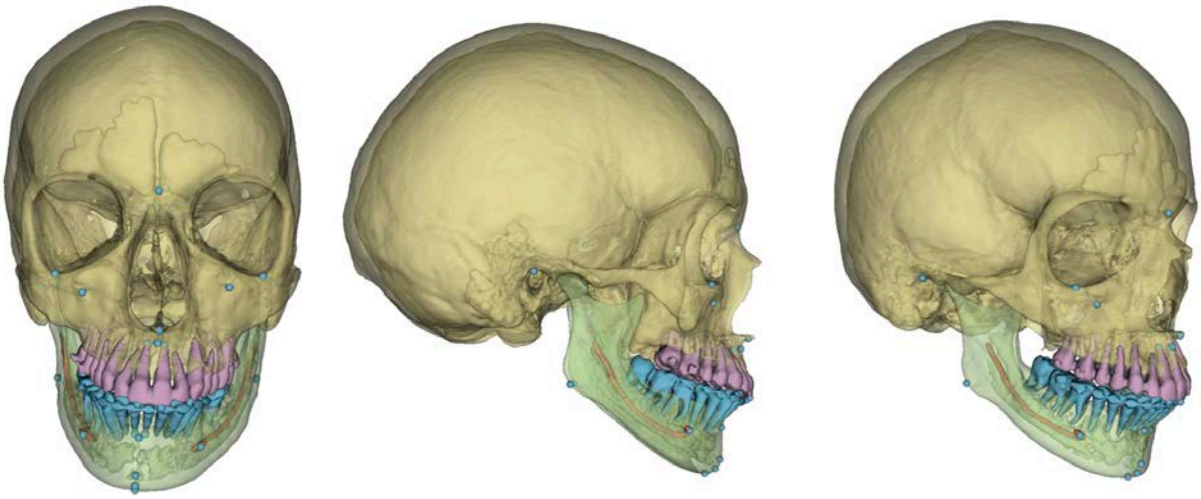


Figure 59 : Cas clinique 3, segmentation automatisée des imageries pré-chirurgicales CT-Scan et placement automatisé des points céphalométriques 3D.

L'analyse géométrique globale (Figure 60) montre qu'il existe une forte asymétrie mandibulaire.

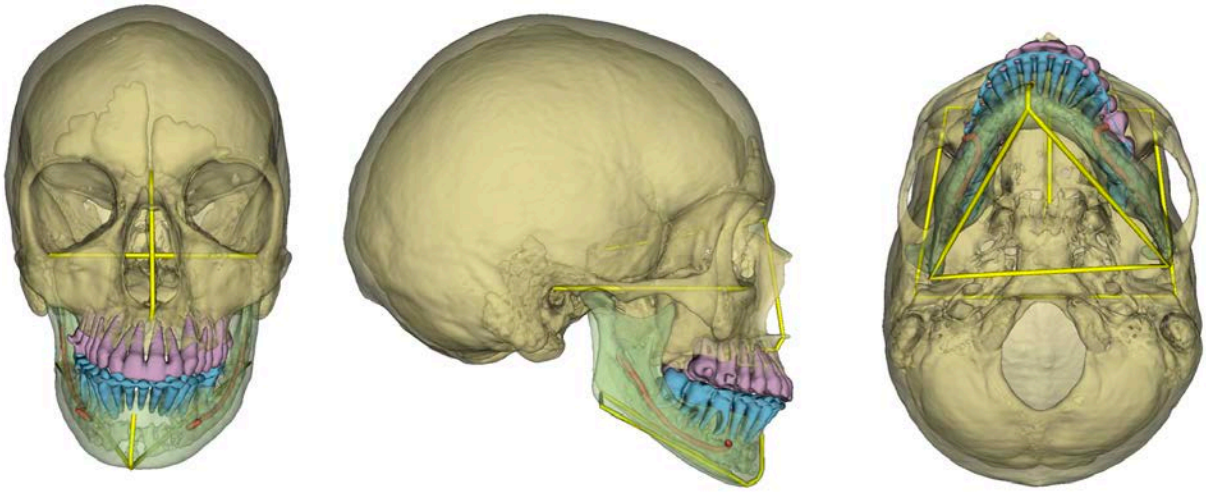


Figure 60 : Cas clinique 3, analyse géométrique globale.

Une analyse géométrique centrée sur le massif crânien supérieur permet de visualiser la position des dents maxillaires par rapport à la base crânienne (Figure 61). Nous mettons ainsi en évidence une asymétrie (plus légère que celle de la mandibule) du maxillaire et une bascule du plan d'occlusion maxillaire.

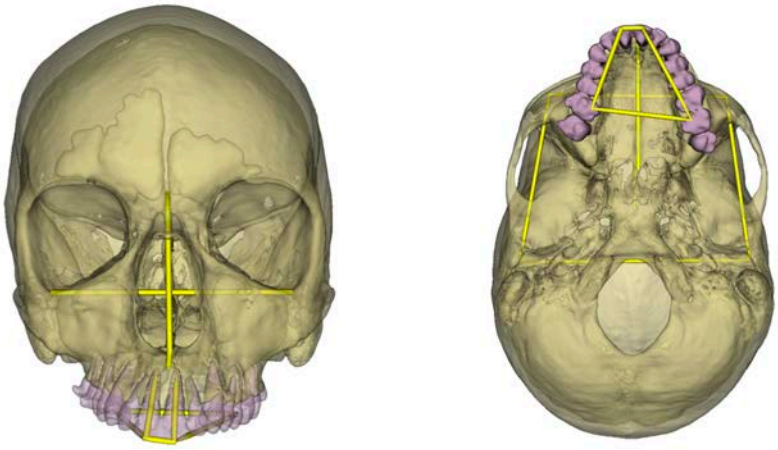


Figure 61 : Cas clinique 3, analyse géométrique du massif cranio-facial et des dents maxillaires. A noter le point 26O incorrectement localisé sur la dent 27.

Une analyse géométrique centrée sur la mandibule montre l'asymétrie flagrante des branches mandibulaires, alors que le corps mandibulaire semble symétrique et les dents mandibulaires correctement situées au sein de celui-ci (Figure 62).

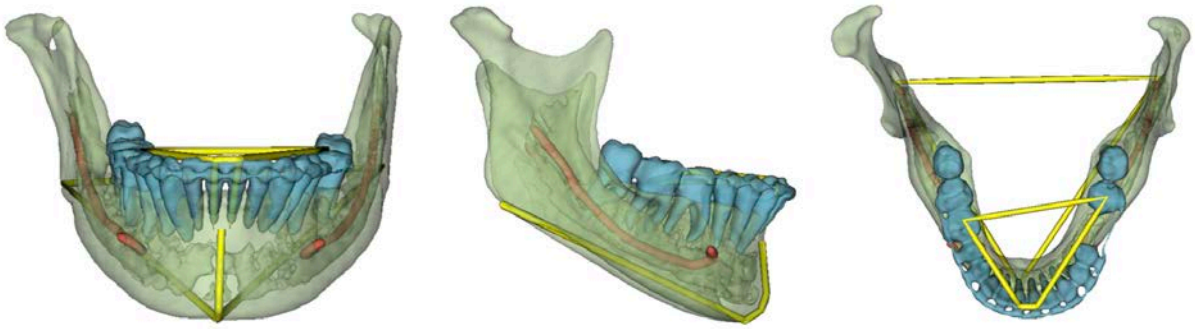


Figure 62 : Cas clinique 3, analyse géométrique de la mandibule et des dents mandibulaires. A noter le point 36O incorrectement localisé sur la dent 37 (la dent 36 étant absente).

Au total, ces éléments vont dans le sens d'une patiente présentant une asymétrie maxillaire et mandibulaire (déviation vers la droite) associée à une bascule du plan d'occlusion (en haut à droite). Ces éléments d'analyse doivent évidemment être confrontés à l'examen clinique pour poser un diagnostic complet et proposer une planification chirurgicale. La planification chirurgicale finalement effectuée recoupe ces éléments : ostéotomie maxillaire pour un léger abaissement à droite et un recentrage du maxillaire avec horizontalisation, ostéotomie mandibulaire de recentrage (Figure 63).

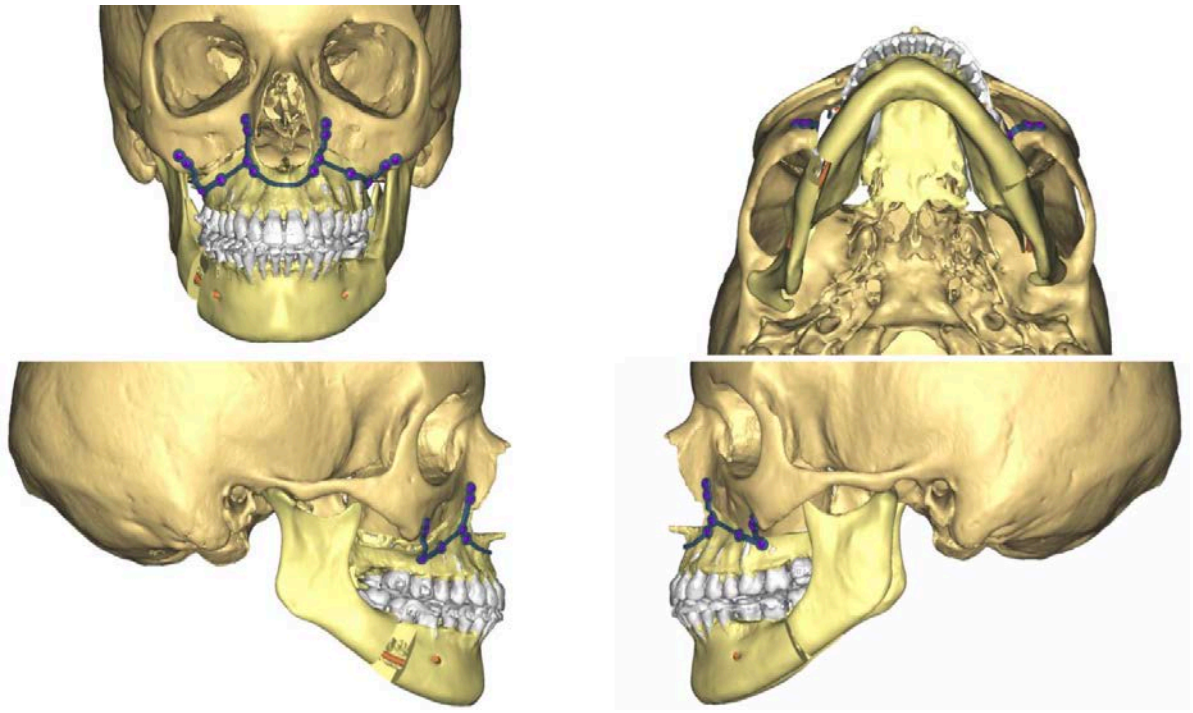


Figure 63 : Cas clinique 3, planification chirurgicale assistée par ordinateur (plaques d'ostéosynthèse sur-mesure).

8.3. Discussion

Nos modèles d'apprentissage profond ont permis de convenablement segmenter et localiser 33 points céphalométriques sur les imageries CT-Scan de ces trois cas cliniques. Ces résultats, obtenus entièrement automatiquement, sont cliniquement exploitables et permettent de visualiser intuitivement l'anatomie des patients en 3D et d'apprécier l'origine des dysmorphies faciales. De tels éléments pourraient être mis à disposition des cliniciens dès le bilan pré-chirurgical, afin de les corréler en direct à l'examen clinique du patient. Cela contribuerait à faciliter l'établissement du diagnostic et la prescription de la planification chirurgicale.

Nous avons proposé ici des analyses géométriques et céphalométriques très simples, permettant avant tout d'illustrer le potentiel clinique de notre travail. Comme détaillé dans le Chapitre 1, les analyses céphalométriques 3D restent complexes à mettre en œuvre et feront l'objet de recherches futures. En particulier, davantage de travaux basés sur la morphométrie sont nécessaires afin d'affiner les diagnostics et proposer des pistes thérapeutiques à partir des emplacements des points céphalométriques (Treil et al. 2020).

Concernant la localisation automatisée des points, deux erreurs au niveau du placement d'un point dentaire (Figure 48, Figure 61, Figure 62) sont à noter, rappelant la nécessité d'une validation des prédictions par des experts humains.

Conclusion générale

L'utilisation clinique des imageries 3D dento-maxillo-faciales se développe fortement, mais le traitement de celles-ci par les cliniciens reste principalement manuel. Le sujet principal de cette thèse était de contribuer à la diffusion clinique de la céphalométrie 3D via l'automatisation de deux étapes, la segmentation et la localisation de points céphalométriques.

La première partie s'est intéressée au contexte et à l'état de l'art du domaine. La céphalométrie 2D, aujourd'hui réalisée en routine, n'est pas suffisante pour l'analyse de certains patients présentant des dysmorphies faciales complexes. Malgré différentes propositions d'analyses 3D, celles-ci ne sont pas réalisées en routine clinique car leur mise en œuvre est complexe : il faut d'abord segmenter les imageries afin de visualiser les modèles 3D, puis ensuite placer les points céphalométriques sur ces modèles. Ces dernières années, plusieurs publications se basant sur des modèles d'apprentissage profond ont rapporté des résultats extrêmement encourageants pour l'automatisation de ces deux tâches. Cependant, ces approches restent préliminaires, les modèles ne sont pas accessibles et nous ne pouvons connaître leur comportement dans un contexte clinique.

La deuxième partie de notre travail a traité de l'automatisation de la segmentation des imageries 3D dento-maxillo-faciales, avec l'entraînement d'un modèle d'apprentissage profond basé sur la méthode librement disponible nnU-Net. La validation de ce modèle a été effectuée sur 153 CT-Scans effectués avant une chirurgie orthognathique, en utilisant des méthodes d'évaluation originales et présentant un intérêt clinique. Ces résultats ont montré la viabilité clinique de notre modèle. Nous avons pu compléter l'évaluation de ce modèle avec un jeu de données de 25 CT-Scans complètement extérieurs à notre base de données, segmentés avec succès.

La troisième partie a concerné le placement des points céphalométriques. Nous avons d'abord mis en place une étude de répétabilité et reproductibilité du placement manuel de 33 points céphalométriques 3D, permettant d'établir l'objectif à atteindre par une approche automatisée. Nous avons ensuite présenté la mise en œuvre et l'entraînement d'un modèle d'apprentissage profond basé sur l'architecture libre SpatialConfiguration-Net. Nous avons évalué ce modèle sur une base de données de 38 CT-Scans pré-chirurgicaux, en utilisant des critères d'évaluation en partie basés sur des mesures céphalométriques ayant un intérêt clinique direct. La méthode a permis de localiser les points

situés sur des surfaces osseuses avec une fiabilité comparable à celle des opérateurs manuels, pendant que la localisation des points sur les surfaces dentaires reste à améliorer.

Enfin, nous avons montré les potentielles applications cliniques de ces deux modèles dans le contexte de la planification chirurgicale par ordinateur. Les trois cas cliniques présentés ont permis d'illustrer la façon dont nos résultats pourraient faciliter ces planifications chirurgicales complexes, pour le bénéfice des patients, des cliniciens et des partenaires industriels.

Au total, ces résultats contribuent à confirmer le potentiel des modèles basés sur l'apprentissage profond pour l'automatisation de la céphalométrie 3D. Ce sont les premiers à évaluer des modèles d'apprentissage profond sur un si grand nombre d'imageries aléatoirement sélectionnées dans une base de données de cas consécutifs d'un service hospitalier. La prochaine étape devra consister en une validation plus large, incluant des données d'autres centres cliniques.

A court terme, ces résultats pourront être utilisés afin de développer une analyse céphalométrique 3D à visée diagnostique. Les annotations manuelles effectuées sur 198 CT-Scans pourraient être complétées des annotations automatiques sur les 255 CT-Scans restant dans notre base de données. Ces données biométriques pourraient être croisées avec les diagnostics cliniques et les planifications chirurgicales afin d'établir des typologies de patients.

A moyen terme, les perspectives sont nombreuses. L'entraînement de nos modèles devra être complété avec des imageries CBCT afin d'automatiser le traitement de ce type d'imageries. Au vu des publications parues sur le sujet, nous pensons que nos approches devraient offrir de bons résultats sur ces données, mais nous n'avons pas pu constituer de base de données de CBCT dans le cadre de ce travail de thèse. Le développement des acquisitions CBCT dites ultra basse dose pourraient augmenter les indications de ces imageries dans le cadre des traitements orthodontiques, et l'automatisation de la céphalométrie 3D à visée orthodontique serait alors d'une grande nécessité. D'autres applications, en médecine légale ou en anthropologie médico-légale, seraient également à développer.

Le succès des potentielles applications cliniques à venir dépendra de la mise en place de méthodes d'évaluation rigoureuses et de la qualité des bases de données utilisées. Dans un contexte industriel extrêmement compétitif, nous encourageons à une collaboration entre les établissements de santé pour constituer de telles bases de données et évaluer prospectivement l'apport de ces approches dans la pratique clinique.

Bibliographie

Alkhayer A, Piffkó J, Lippold C, Segatto E. 2020. Accuracy of virtual planning in orthognathic surgery: a systematic review. Head Face Med. 16(1):34.

American Academy of Oral and Maxillofacial Radiology. 2013. Clinical recommendations regarding use of cone beam computed tomography in orthodontics. Position statement by the American Academy of Oral and Maxillofacial Radiology. Oral Surg Oral Med Oral Pathol Oral Radiol. 116(2):238–257.

Bermejo E, Taniguchi K, Ogawa Y, Martos R, Valsecchi A, Mesejo P, Ibáñez O, Imaizumi K. 2021. Automatic landmark annotation in 3D surface scans of skulls: Methodological proposal and reliability study. Comput Methods Programs Biomed. 210:106380.

Bichu YM, Hansa I, Bichu AY, Premjani P, Flores-Mir C, Vaid NR. 2021. Applications of artificial intelligence and machine learning in orthodontics: a scoping review. Prog Orthod. 22(1):18.

Bland JM, Altman DG. 1996. Statistics Notes: Measurement error. BMJ. 313(7059):744.

Borzabadi-Farahani A, Eslamipour F, Shahmoradi M. 2016. Functional needs of subjects with dentofacial deformities: A study using the index of orthognathic functional treatment need (IOFTN). J Plast Reconstr Aesthet Surg. 69(6):796–801.

van Bunningen RH, Dijkstra PU, Dieters A, van der Meer WJ, Kuijpers-Jagtman AM, Ren Y. 2022. Precision of orthodontic cephalometric measurements on ultra low dose-low dose CBCT reconstructed cephalograms. Clin Oral Investig. 26(2):1543–1550.

Cattaneo PM, Yung AKC, Holm A, Mashaly OM, Cornelis MA. 2019. 3D landmarks of Craniofacial Imaging and subsequent considerations on superimpositions in orthodontics—The Aarhus perspective. Orthod Craniofac Res. 22(S1):21–29.

Chen A, Dawant B. 2016 Feb 2. A Multi-atlas Approach for the Automatic Segmentation of Multiple Structures in Head and Neck CT Images. MIDAS J. [accessed 2022 Mar 1]. https://www.midasjournal.org/browse/publication/964.

Chen R, Ma Y, Chen N, Liu L, Cui Z, Lin Y, Wang W. 2022. Structure-Aware Long Short-Term Memory Network for 3D Cephalometric Landmark Detection. IEEE Trans Med Imaging.:1–1.

Chen R, Ma Y, Liu L, Chen N, Cui Z, Wei G, Wang W. 2022. Semi-supervised anatomical landmark detection via shape-regulated self-training. Neurocomputing. 471:335–345.

Chen X, Lian C, Deng HH, Kuang T, Lin H-Y, Xiao D, Gateno J, Shen D, Xia JJ, Yap P-T. 2021. Fast and Accurate Craniomaxillofacial Landmark Detection via 3D Faster R-CNN. IEEE Trans Med Imaging. 40(12):3867–3878.

Chien P, Parks E, Eraso F, Hartsfield J, Roberts W, Ofner S. 2009. Comparison of reliability in anatomical landmark identification using two-dimensional digital cephalometrics and three-dimensional cone beam computed tomography *in vivo*. Dentomaxillofacial Radiol. 38(5):262–273.

Chung M, Lee M, Hong J, Park S, Lee Jusang, Lee Jingyu, Yang I-H, Lee Jeongjin, Shin Y-G. 2020. Pose-aware instance segmentation framework from cone beam CT images for tooth segmentation. Comput Biol Med. 120:103720.

Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. Vol. 9901. Cham: Springer International Publishing. (Lecture Notes in Computer Science). p. 424–432. [accessed 2021 Feb 11]. http://link.springer.com/10.1007/978-3-319-46723-8_49.

Codari M, Caffini M, Tartaglia GM, Sforza C, Baselli G. 2017. Computer-aided cephalometric landmark annotation for CBCT data. Int J Comput Assist Radiol Surg. 12(1):113–121.

Cui Z, Li C, Wang W. 2019. ToothNet: Automatic Tooth Instance Segmentation and Identification From

Cone Beam CT Images. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE. p. 6361–6370. [accessed 2021 Feb 9]. https://ieeexplore.ieee.org/document/8954147/.

Dalal N, Triggs B. 2005. Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. p. 886–893 vol. 1.

Damas S, Cordón O, Ibáñez O. 2018. Handbook on craniofacial superimposition: the meprocs project. New York, NY: Springer Berlin Heidelberg.

Di Angelo L, Di Stefano P, Governi L, Marzola A, Volpe Y. 2019. A Robust and Automatic Method for the Best Symmetry Plane Detection of Craniofacial Skeletons. Symmetry. 11(2):245.

Donatelli RE, Lee S-J. 2013a. How to report reliability in orthodontic research: Part 1. Am J Orthod Dentofacial Orthop. 144(1):156–161.

Donatelli RE, Lee S-J. 2013b. How to report reliability in orthodontic research: Part 2. Am J Orthod Dentofacial Orthop. 144(2):315–318.

Dot G, Rafflenbeul F, Arbotto M, Gajny L, Rouch P, Schouman T. 2020. Accuracy and reliability of automatic three-dimensional cephalometric landmarking. Int J Oral Maxillofac Surg. 49(10):1367–1378.

Dot G, Rafflenbeul F, Kerbrat A, Rouch P, Gajny L, Schouman T. 2021. Three-Dimensional Cephalometric Landmarking and Frankfort Horizontal Plane Construction: Reproducibility of Conventional and Novel Landmarks. J Clin Med. 10(22):5303.

Dot G, Schouman T, Chang S, Rafflenbeul F, Kerbrat A, Rouch P, Gajny L. 2022 Aug 18. Automatic 3-Dimensional Cephalometric Landmarking via Deep Learning. J Dent Res.:00220345221112333.

Dot G, Schouman T, Dubois G, Rouch P, Gajny L. 2022. Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnU-Net framework. Eur Radiol. 32(6):3639–3648.

Egger J, Pfarrkirchner B, Gsaxner C, Lindner L, Schmalstieg D, Wallner J. 2018. Fully Convolutional Mandible Segmentation on a valid Ground- Truth Dataset. Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf. 2018:656–660.

Falk T, Mai D, Bensch R, Çiçek Ö, Abdulkadir A, Marrakchi Y, Böhm A, Deubner J, Jäckel Z, Seiwald K, et al. 2019. U-Net: deep learning for cell counting, detection, and morphometry. Nat Methods. 16(1):67–70.

Faure J, Baron P, Treil J. 2005. Analyse céphalométrique tridimentionnelle : diagnostic des dysmorphies antéropostérieures et verticales. Orthod Fr. 76(2):91–110.

Faure J, Oueiss A, Treil J, Chen S, Wong V, Inglese J-M. 2016. Céphalométrie 3D et intelligence artificielle. Rev Orthopédie Dento-Faciale. 50(3):315–334.

Fontenele RC, Gerhardt M do N, Pinto JC, Van Gerven A, Willems H, Jacobs R, Freitas DQ. 2022. Influence of dental fillings and tooth type on the performance of a novel artificial intelligence-driven tool for automatic tooth segmentation on CBCT images – A validation study. J Dent. 119:104069.

Frongia G, Grazia Piancino M, Adriano Bracco A, Crincoli V, Lorenzo Debernardi C, Bracco P. 2012. Assessment of the Reliability and Repeatability of Landmarks Using 3-D Cephalometric Software. CRANIO®. 30(4):255–263.

Fryback DG, Thornbury JR. 1991. The efficacy of diagnostic imaging. Med Decis Mak Int J Soc Med Decis Mak. 11(2):88–94.

Gateno J, Xia JJ, Teichgraeber JF. 2011. New 3-Dimensional Cephalometric Analysis for Orthognathic Surgery. J Oral Maxillofac Surg. 69(3):606–622.

Gollmer ST, Buzug TM. 2012. Fully automatic shape constrained mandible segmentation from cone-beam CT data. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI). p. 1272–1275.

Granata S, Giberti L, Vigolo P, Stellini E, Fiore AD. 2020. Incorporating a facial scanner into the digital workflow: A dental technique. J Prosthet Dent. 123(6):781–785.

Gribel BF, Gribel MN, Frazão DC, McNamara JA, Manzi FR. 2011. Accuracy and reliability of craniometric

measurements on lateral cephalometry and 3D measurements on CBCT scans. Angle Orthod. 81(1):26–35.
Gupta A, Kharbanda OP, Sardana V, Balachandran R, Sardana HK. 2015. A knowledge-based algorithm for automatic detection of cephalometric landmarks on CBCT images. Int J Comput Assist Radiol Surg. 10(11):1737–1752.
Gupta A, Kharbanda OP, Sardana V, Balachandran R, Sardana HK. 2016. Accuracy of 3D cephalometric measurements based on an automatic knowledge-based landmark detection algorithm. Int J Comput Assist Radiol Surg. 11(7):1297–1309.
Hao J, Liao W, Zhang YL, Peng J, Zhao Z, Chen Z, Zhou BW, Feng Y, Fang B, Liu ZZ, et al. 2021 Nov 1. Toward Clinically Applicable 3-Dimensional Tooth Segmentation via Deep Learning. J Dent Res.:002203452110404.
Hassan B, Nijkamp P, Verheij H, Tairie J, Vink C, van der Stelt P, van Beek H. 2013. Precision of identifying cephalometric landmarks with cone beam computed tomography in vivo. Eur J Orthod. 35(1):38–44.
Ho C-T, Lin H-H, Liou EJW, Lo L-J. 2017. Three-dimensional surgical simulation improves the planning for correction of facial prognathism and asymmetry: A qualitative and quantitative study. Sci Rep. 7(1):40423.
Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 18(2):203–211.
ISO 5725-2:2019. Accuracy (trueness and precision) of measurement methods and results.
Jaskari J, Sahlsten J, Järnstedt J, Mehtonen H, Karhu K, Sundqvist O, Hietanen A, Varjonen V, Mattila V, Kaski K. 2020. Deep Learning Method for Mandibular Canal Segmentation in Dental Cone Beam Computed Tomography Volumes. Sci Rep. 10(1):5842.
de Jong MA, Gül A, de Gijt JP, Koudstaal MJ, Kayser M, Wolvius EB, Böhringer S. 2018. Automated human skull landmarking with 2D Gabor wavelets. Phys Med Biol. 63(10):105011.
Juerchott A, Freudlsperger C, Weber D, Jende JME, Saleem MA, Zingler S, Lux CJ, Bendszus M, Heiland S, Hilgenfeld T. 2020. In vivo comparison of MRI- and CBCT-based 3D cephalometric analysis: beginning of a non-ionizing diagnostic era in craniomaxillofacial imaging? Eur Radiol. 30(3):1488–1497.
Kang SH, Jeon K, Kang S-H, Lee S-H. 2021. 3D cephalometric landmark detection by multiple stage deep reinforcement learning. Sci Rep. 11(1):17509.
Kapila SD, Nervina JM. 2015. CBCT in orthodontics: assessment of treatment outcomes and indications for its use. Dentomaxillofacial Radiol. 44(1):20140282.
Krizhevsky A, Sutskever I, Hinton GE. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems. Vol. 25. Curran Associates, Inc. [accessed 2022 Mar 2]. https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html.
Kumar V, Ludlow J, Soares Cevidanes LH, Mol A. 2008. In Vivo Comparison of Conventional and Cone Beam CT Synthesized Cephalograms. Angle Orthod. 78(5):873–879.
Kwak GH, Kwak E-J, Song JM, Park HR, Jung Y-H, Cho B-H, Hui P, Hwang JJ. 2020. Automatic mandibular canal detection using a deep convolutional neural network. Sci Rep. 10(1):5711.
Lagravère MO, Low C, Flores-Mir C, Chung R, Carey JP, Heo G, Major PW. 2010. Intraexaminer and interexaminer reliabilities of landmark identification on digitized lateral cephalograms and formatted 3-dimensional cone-beam computerized tomography images. Am J Orthod Dentofacial Orthop. 137(5):598–604.
Lahoud P, Diels S, Niclaes L, Van Aelst S, Willems H, Van Gerven A, Quirynen M, Jacobs R. 2022. Development and validation of a novel artificial intelligence driven tool for accurate mandibular canal segmentation on CBCT. J Dent. 116:103891.
Landi F, O'Higgins P. 2019. Applying Geometric Morphometrics to Digital Reconstruction and Anatomical Investigation. In: Rea PM, editor. Biomedical Visualisation: Volume 4. Cham: Springer International Publishing. (Advances in Experimental Medicine and Biology). p. 55–71.

[accessed 2022 Mar 18]. https://doi.org/10.1007/978-3-030-24281-7_6.
Lang Y, Lian C, Xiao D, Deng H, Yuan P, Gateno J, Shen SGF, Alfi DM, Yap P-T, Xia JJ, et al. 2020. Automatic Localization of Landmarks in Craniomaxillofacial CBCT Images Using a Local Attention-Based Graph Convolution Network. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Vol. 12264. Cham: Springer International Publishing. (Lecture Notes in Computer Science). p. 817–826. [accessed 2021 Dec 22]. https://link.springer.com/10.1007/978-3-030-59719-1_79.
LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature. 521(7553):436–444.
Lee S-H, Kil T-J, Park K-R, Kim BC, Kim J-G, Piao Z, Corre P. 2014. Three-dimensional architectural and structural analysis--a transition in concept and design from Delaire's cephalometric analysis. Int J Oral Maxillofac Surg. 43(9):1154–1160.
Lee SM, Kim HP, Jeon K, Lee S-H, Seo JK. 2019. Automatic 3D cephalometric annotation system using shadowed 2D image-based machine learning. Phys Med Biol. 64(5):055002.
Leonardi R, Giordano D, Maiorana F, Spampinato C. 2008. Automatic Cephalometric Analysis: A Systematic Review. Angle Orthod. 78(1):145–151.
Leonardi R, Giudice AL, Isola G, Spampinato C. 2021. Deep learning and computer vision: Two promising pillars, powering the future in orthodontics. Semin Orthod. 27(2):62–68.
Lian C, Wang F, Deng HH, Wang L, Xiao D, Kuang T, Lin H-Y, Gateno J, Shen SGF, Yap P-T, et al. 2020. Multi-task Dynamic Transformer Network for Concurrent Bone Segmentation and Large-Scale Landmark Localization with Dental CBCT. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Vol. 12264. Cham: Springer International Publishing. (Lecture Notes in Computer Science). p. 807–816. [accessed 2021 Feb 9]. http://link.springer.com/10.1007/978-3-030-59719-1_78.
Lim B-D, Choi D-S, Jang I, Cha B-K. 2019. Application of the foramina of the trigeminal nerve as landmarks for analysis of craniofacial morphology. Korean J Orthod. 49(5):326.
Lin H-H, Chuang Y-F, Weng J-L, Lo L-J. 2015. Comparative Validity and Reproducibility Study of Various Landmark-Oriented Reference Planes in 3-Dimensional Computed Tomographic Analysis for Patients Receiving Orthognathic Surgery. Elsalanty M, editor. PLOS ONE. 10(2):e0117604.
Lindner C, Wang C-W, Huang C-T, Li C-H, Chang S-W, Cootes TF. 2016. Fully automatic system for accurate localisation and analysis of cephalometric landmarks in lateral cephalograms. Sci Rep. 6:33581.
Lisboa C de O, Masterson D, Motta AFJ, Motta AT. 2015. Reliability and reproducibility of three-dimensional cephalometric landmarks using CBCT: a systematic review. J Appl Oral Sci. 23(2):112–119.
Liu Q, Deng H, Lian C, Chen Xiaoyang, Xiao D, Ma L, Chen Xu, Kuang T, Gateno J, Yap P-T, et al. 2021. SkullEngine: A Multi-stage CNN Framework for Collaborative CBCT Image Segmentation and Landmark Detection. In: Lian C, Cao X, Rekik I, Xu X, Yan P, editors. Machine Learning in Medical Imaging. Vol. 12966. Cham: Springer International Publishing. (Lecture Notes in Computer Science). p. 606–614. [accessed 2021 Dec 22]. https://link.springer.com/10.1007/978-3-030-87589-3_62.
Lonic D, Sundoro A, Lin H-H, Lin P-J, Lo L-J. 2017. Selection of a horizontal reference plane in 3D evaluation: Identifying facial asymmetry and occlusal cant in orthognathic surgery planning. Sci Rep. 7(1):2157.
Ludlow JB, Gubler M, Cevidanes L, Mol A. 2009. Precision of cephalometric landmark identification: Cone-beam computed tomography vs conventional cephalometric views. Am J Orthod Dentofacial Orthop. 136(3):312.e1-312.e10.
Ma Q, Kobayashi E, Fan B, Nakagawa K, Sakuma I, Masamune K, Suenaga H. 2020. Automatic 3D landmarking model using patch-based deep neural networks for CT image of oral and maxillofacial surgery. Int J Med Robot. 16(3). [accessed 2021 Dec 22]. https://onlinelibrary.wiley.com/doi/10.1002/rcs.2093.

McAlinden C, Khadka J, Pesudovs K. 2015. Precision (repeatability and reproducibility) studies and sample-size calculation. J Cataract Refract Surg. 41(12):2598–2604.

Minnema J, Eijnatten M, Hendriksen AA, Liberton N, Pelt DM, Batenburg KJ, Forouzanfar T, Wolff J. 2019. Segmentation of dental cone-beam CT scans affected by metal artifacts using a mixed-scale dense convolutional neural network. Med Phys. 46(11):5027–5035.

Montúfar J, Romero M, Scougall-Vilchis RJ. 2018. Hybrid approach for automatic cephalometric landmark annotation on cone-beam computed tomography volumes. Am J Orthod Dentofac Orthop Off Publ Am Assoc Orthod Its Const Soc Am Board Orthod. 154(1):140–150.

Mörch CM, Atsu S, Cai W, Li X, Madathil SA, Liu X, Mai V, Tamimi F, Dilhac MA, Ducret M. 2021. Artificial Intelligence and Ethics in Dentistry: A Scoping Review. J Dent Res. 100(13):1452–1460.

Moshiri M, Scarfe WC, Hilgers ML, Scheetz JP, Silveira AM, Farman AG. 2007. Accuracy of linear measurements from imaging plate and lateral cephalometric images derived from cone-beam computed tomography. Am J Orthod Dentofacial Orthop. 132(4):550–560.

Muehlematter UJ, Daniore P, Vokinger KN. 2021. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Lancet Digit Health. 3(3):e195–e203.

Murabito F, Palazzo S, Salanitri FP, Rundo F, Bagci U, Giordano D, Leonardi R, Spampinato C. 2021. Deep Recurrent-Convolutional Model for Automated Segmentation of Craniomaxillofacial CT Scans. In: 2020 25th International Conference on Pattern Recognition (ICPR). Milan, Italy: IEEE. p. 9062–9067. [accessed 2021 Jun 1]. https://ieeexplore.ieee.org/document/9413084/.

Naji P, Alsufyani NA, Lagravère MO. 2014. Reliability of anatomic structures as landmarks in three-dimensional cephalometric analysis using CBCT. Angle Orthod. 84(5):762–772.

Neelapu BC, Kharbanda OP, Sardana V, Gupta A, Vasamsetti S, Balachandran R, Sardana HK. 2018. Automatic localization of three-dimensional cephalometric landmarks on CBCT images by extracting symmetry features of the skull. Dento Maxillo Facial Radiol. 47(2):20170054.

da Neiva MB, Soares ÁC, de Oliveira Lisboa C, de Vasconcellos Vilella O, Motta AT. 2015. Evaluation of cephalometric landmark identification on CBCT multiplanar and 3D reconstructions. Angle Orthod. 85(1):11–17.

Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, Patel Y, Meyer C, Askham H, Romera-Paredes B, et al. 2021. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. J Med Internet Res. 23(7):e26151.

Noori SMR, Farnia P, Bayat M, Bahrami N, Shakourirad A, Ahmadian A. 2020. Automatic detection of symmetry plane for computer-aided surgical simulation in craniomaxillofacial surgery. Phys Eng Sci Med. 43(3):1087–1099.

Oh S, Ahn J, Nam K-U, Paeng J-Y, Hong J. 2013. Frankfort horizontal plane is an appropriate three-dimensinal reference in the evaluation of clinical and skeletal cant. J Korean Assoc Oral Maxillofac Surg. 39(2):71.

de Oliveira AEF, Cevidanes LHS, Phillips C, Motta A, Burke B, Tyndall D. 2009. Observer reliability of three-dimensional cephalometric landmark identification on cone-beam computerized tomography. Oral Surg Oral Med Oral Pathol Oral Radiol Endodontology. 107(2):256–265.

Olszewski R, Cosnard G, Macq B, Mahy P, Reychler H. 2006. 3D CT-based cephalometric analysis: 3D cephalometric theoretical concept and software. Neuroradiology. 48(11):853–862.

Olszewski R, Tanesy O, Cosnard G, Zech F, Reychler H. 2010. Reproducibility of osseous landmarks used for computed tomography based three-dimensional cephalometric analyses. J Cranio-Maxillofac Surg. 38(3):214–221.

O'Neil AQ, Kascenas A, Henry J, Wyeth D, Shepherd M, Beveridge E, Clunie L, Sansom C, Šeduikytė E, Muir K, et al. 2019. Attaining Human-Level Performance with Atlas Location Autocontext for Anatomical Landmark Detection in 3D CT Data. In: Leal-Taixé L, Roth S, editors. Computer Vision – ECCV 2018 Workshops. Springer International Publishing. p. 470–484.

Oueiss A, Treil J, Faure J. 2020. Biométrie cranio-faciale 3D: analyse complète d'un cas de classe II « limite chirurgicale ». Orthod Fr. 91(1):115–128.

Patel A, Otterburn D, Saadeh P, Levine J, Hirsch DL. 2011. 3D Volume Assessment Techniques and Computer-Aided Design and Manufacturing for Preoperative Fabrication of Implants in Head and Neck Reconstruction. Facial Plast Surg Clin N Am. 19(4):683–709.

Pauwels R, Araki K, Siewerdsen JH, Thongvigitmanee SS. 2015. Technical aspects of dental CBCT: state of the art. Dentomaxillofacial Radiol. 44(1):20140224.

Payer C, Štern D, Bischof H, Urschler M. 2019. Integrating spatial configuration into heatmap regression based CNNs for landmark localization. Med Image Anal. 54:207–219.

Pietzka S, Wilde F, Schramm A, Mascha F. 2019. Navigated orbital floor reconstruction with cad/cam guide and patient-specific implant. Int J Oral Maxillofac Surg. 48:28.

Pinheiro M, Ma X, Fagan MJ, McIntyre GT, Lin P, Sivamurthy G, Mossey PA. 2019. A 3D cephalometric protocol for the accurate quantification of the craniofacial symmetry and facial growth. J Biol Eng. 13(1):42.

Pittayapat P, Jacobs R, Bornstein MM, Odri GA, Lambrichts I, Willems G, Politis C, Olszewski R. 2018. Three-dimensional Frankfort horizontal plane for 3D cephalometry: a comparative assessment of conventional versus novel landmarks and horizontal planes. Eur J Orthod. 40(3):239–248.

Pittayapat P, Limchaichana-Bolstad N, Willems G, Jacobs R. 2014. Three-dimensional cephalometric analysis in orthodontics: a systematic review. Orthod Craniofac Res. 17(2):69–91.

Proffit WR, Sarver DM, Fields HW. 2018. Orthodontic Diagnosis: The Problem-Oriented Approach. In: Contemporary orthodontics. 6th edition. Philadelphia, IL: Elsevier. p. 140–207.

Qiu B, Guo J, Kraeima J, Glas HH, Borra RJH, Witjes MJH, van Ooijen PMA. 2019. Automatic segmentation of the mandible from computed tomography scans for 3D virtual surgical planning using the convolutional neural network. Phys Med Biol. 64(17):175020.

Qiu B, Guo J, Kraeima J, Glas HH, Zhang W, Borra RJH, Witjes MJH, van Ooijen PMA. 2021. Recurrent Convolutional Neural Networks for 3D Mandible Segmentation in Computed Tomography. J Pers Med. 11(6):492.

Quast A, Santander P, Witt D, Damm A, Moser N, Schliephake H, Meyer-Marcotty P. 2019. Traditional face-bow transfer versus three-dimensional virtual reconstruction in orthognathic surgery. Int J Oral Maxillofac Surg. 48(3):347–354.

Reinke A, Eisenmann M, Tizabi MD, Sudre CH, Rädsch T, Antonelli M, Arbel T, Bakas S, Cardoso MJ, Cheplygina V, et al. 2021 Apr 13. Common Limitations of Image Processing Metrics: A Picture Story. ArXiv210405642 Cs Eess. [accessed 2021 May 31]. http://arxiv.org/abs/2104.05642.

Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. 2009. Assessing methodological quality. Cochrane Handb Syst Rev Diagn Test Accuracy Version. 1(0).

Ronneberger O, Fischer P, Brox T. 2015 May 18. U-Net: Convolutional Networks for Biomedical Image Segmentation. ArXiv150504597 Cs. [accessed 2021 Feb 9]. http://arxiv.org/abs/1505.04597.

Ruellas AC de O, Tonello C, Gomes LR, Yatabe MS, Macron L, Lopinto J, Goncalves JR, Garib Carreira DG, Alonso N, Souki BQ, et al. 2016. Common 3-dimensional coordinate system for assessment of directional changes. Am J Orthod Dentofacial Orthop. 149(5):645–656.

Saha S. 2018 Dec 17. A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way. Medium. [accessed 2022 Mar 3]. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

Salameh J-P, Bossuyt PM, McGrath TA, Thombs BD, Hyde CJ, Macaskill P, Deeks JJ, Leeflang M, Korevaar DA, Whiting P, et al. 2020. Preferred reporting items for systematic review and meta-analysis of diagnostic test accuracy studies (PRISMA-DTA): explanation, elaboration, and checklist. BMJ. 370. [accessed 2020 Sep 17]. https://www.bmj.com/content/370/bmj.m2632.

Sam A, Currie K, Oh H, Flores-Mir C, Lagravére-Vich M. 2018. Reliability of different three-dimensional cephalometric landmarks in cone-beam computed tomography: A systematic review. Angle Orthod. 89(2):317–332.

Santos RMG dos, De Martino JM, Haiter Neto F, Passeri LA. 2017. Influence of different setups of the Frankfort horizontal plane on 3-dimensional cephalometric measurements. Am J Orthod Dentofacial Orthop. 152(2):242–249.

Schlicher W, Nielsen I, Huang JC, Maki K, Hatcher DC, Miller AJ. 2012. Consistency and precision of

landmark identification in three-dimensional cone beam computed tomography scans. Eur J Orthod. 34(3):263–275.
Schouman T, Murcier G, Goudot P. 2015. The key to accuracy of zygoma repositioning: Suitability of the SynpliciTi customized guide-plates. J Cranio-Maxillofac Surg. 43(10):1942–1947.
Schwendicke F, Chaurasia A, Arsiwala L, Lee J-H, Elhennawy K, Jost-Brinkmann P-G, Demarco F, Krois J. 2021. Deep learning for cephalometric landmark detection: systematic review and meta-analysis. Clin Oral Investig. 25(7):4299–4309.
Schwendicke F, Golla T, Dreher M, Krois J. 2019. Convolutional neural networks for dental image diagnostics: A scoping review. J Dent. 91:103226.
Schwendicke F, Samek W, Krois J. 2020. Artificial Intelligence in Dentistry: Chances and Challenges. J Dent Res. 99(7):769–774.
Schwendicke F, Singh T, Lee J-H, Gaudin R, Chaurasia A, Wiegand T, Uribe S, Krois J. 2021. Artificial intelligence in dental research: Checklist for authors, reviewers, readers. J Dent. 107:103610.
SEDENTEXCT project. 2012. Cone Beam CT for dental and maxillofacial radiology (evidence based guidelines). European Commission.
Sekuboyina A, Husseini ME, Bayat A, Löffler M, Liebl H, Li H, Tetteh G, Kukačka J, Payer C, Štern D, et al. 2021. VerSe: A Vertebrae labelling and segmentation benchmark for multi-detector CT images. Med Image Anal. 73:102166.
Shahen S, Lagravère MO, Carrino G, Fahim F, Abdelsalam R, Flores-Mir C, Perillo L. 2018. United Reference Method for three-dimensional treatment evaluation. Prog Orthod. 19(1):47.
Shahidi S, Bahrampour E, Soltanimehr E, Zamani A, Oshagh M, Moattari M, Mehdizadeh A. 2014. The accuracy of a designed software for automated localization of craniofacial landmarks on CBCT images. BMC Med Imaging. 14:32.
Shorten C, Khoshgoftaar TM. 2019. A survey on Image Data Augmentation for Deep Learning. J Big Data. 6(1):60.
Smektała T, Jędrzejewski M, Szyndel J, Sporniak-Tutak K, Olszewski R. 2014. Experimental and clinical assessment of three-dimensional cephalometry: A systematic review. J Cranio-Maxillofac Surg. 42(8):1795–1801.
Swennen GRJ, Schutyser F, Barth E-L, De Groeve P, De Mey A. 2006. A New Method of 3-D Cephalometry Part I: The Anatomic Cartesian 3-D Reference System. J Craniofac Surg. 17(2):314–325.
Swennen GRJ, Schutyser F, Hausamen J-E. 2006. Three-Dimensional Cephalometry: A Color Atlas and Manual. Berlin Heidelberg: Springer-Verlag. [accessed 2019 Mar 22]. https://www.springer.com/gp/book/9783540254409.
Thurzo A, Kosnáčová HS, Kurilová V, Kosmeľ S, Beňuš R, Moravanský N, Kováč P, Kuracinová KM, Palkovič M, Varga I. 2021. Use of Advanced Artificial Intelligence in Forensic Medicine, Forensic Anthropology and Clinical Anatomy. Healthcare. 9(11):1545.
Titiz I, Laubinger M, Keller T, Hertrich K, Hirschfelder U. 2012. Repeatability and reproducibility of landmarks--a three-dimensional computed tomography study. Eur J Orthod. 34(3):276–286.
Torosdagli N, Liberton DK, Verma P, Sincan M, Lee J, Pattanaik S, Bagci U. 2017. Robust and fully automated segmentation of mandible from CT scans. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). Melbourne, Australia: IEEE. p. 1209–1212. [accessed 2021 Feb 9]. http://ieeexplore.ieee.org/document/7950734/.
Torosdagli N, Liberton DK, Verma P, Sincan M, Lee JS, Bagci U. 2019. Deep Geodesic Learning for Segmentation and Anatomical Landmarking. IEEE Trans Med Imaging. 38(4):919–931.
Treil J, Oueiss A, Faure J. 2020. Biométrie cranio-faciale 3D: analyse statistique des dysmorphies de classe II. Orthod Fr. 91(1):101–114.
Vandaele R, Aceto J, Muller M, Péronnet F, Debat V, Wang C-W, Huang C-T, Jodogne S, Martinive P, Geurts P, et al. 2018. Landmark detection in 2D bioimages for geometric morphometrics: a multi-resolution tree-based approach. Sci Rep. 8(1). [accessed 2019 Mar 15]. http://www.nature.com/articles/s41598-017-18993-5.
Verhelst P-J, Smolders A, Beznik T, Meewis J, Vandemeulebroucke A, Shaheen E, Van Gerven A, Willems

H, Politis C, Jacobs R. 2021. Layered deep learning for automatic mandibular segmentation in cone-beam computed tomography. J Dent. 114:103786.
Viola P, Jones M. 2001. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Vol. 1. p. I–I.
Wallner J, Mischak I, Jan Egger. 2019. Computed tomography data collection of the complete human mandible and valid clinical ground truth models. Sci Data. 6(1):190003.
Wallner J, Schwaiger M, Hochegger K, Gsaxner C, Zemann W, Egger J. 2019. A review on multiplatform evaluations of semi-automatic open-source based image segmentation for cranio-maxillofacial surgery. Comput Methods Programs Biomed. 182:105102.
Wang C-W, Huang C-T, Lee J-H, Li C-H, Chang S-W, Siao M-J, Lai T-M, Ibragimov B, Vrtovec T, Ronneberger O, et al. 2016. A benchmark for comparison of dental radiography analysis algorithms. Med Image Anal. 31:63–76.
Wang H, Minnema J, Batenburg KJ, Forouzanfar T, Hu FJ, Wu G. 2021. Multiclass CBCT Image Segmentation for Orthodontics with Deep Learning. J Dent Res. 100(9):943–949.
Whiting PF. 2011. QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. Ann Intern Med. 155(8):529.
Xia JJ, Gateno J, Teichgraeber JF. 2009. New Clinical Protocol to Evaluate Craniomaxillofacial Deformity and Plan Surgical Correction. J Oral Maxillofac Surg. 67(10):2093–2106.
Yatabe M, Gomes L, Ruellas AC, Lopinto J, Macron L, Paniagua B, Budin F, Prieto JC, Ioshida M, Cevidanes L. 2019. Challenges in measuring angles between craniofacial structures. J Appl Oral Sci. 27. [accessed 2022 Mar 11]. http://www.scielo.br/j/jaos/a/PZtM9DGFFWB3yQW53QcYW8n/?lang=en.
Yun HS, Jang TJ, Lee SM, Lee S-H, Seo JK. 2020. Learning-based local-to-global landmark annotation for automatic 3D cephalometry. Phys Med Biol. 65(8):085018.
Zamora N, Llamas Jm, Cibrian R, Gandia J, Paredes V. 2012. A study on the reproducibility of cephalometric landmarks when undertaking a three-dimensional (3D) cephalometric analysis. Med Oral Patol Oral Cirugia Bucal.:e678–e688.
Zhang J, Gao Y, Wang L, Tang Z, Xia JJ, Shen D. 2016. Automatic Craniomaxillofacial Landmark Digitization via Segmentation-Guided Partially-Joint Regression Forest Model and Multiscale Statistical Features. IEEE Trans Biomed Eng. 63(9):1820–1829.
Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen SG-F, Tang Z, Chen K-C, Xia JJ, et al. 2017. Joint Craniomaxillofacial Bone Segmentation and Landmark Digitization by Context-Guided Fully Convolutional Networks. Med Image Comput Comput-Assist Interv MICCAI Int Conf Med Image Comput Comput-Assist Interv. 10434:720–728.
Zhang J, Liu M, Wang L, Chen S, Yuan P, Li J, Shen SG-F, Tang Z, Chen K-C, Xia JJ, et al. 2020. Context-guided fully convolutional networks for joint craniomaxillofacial bone segmentation and landmark digitization. Med Image Anal. 60:101621.
Zhao M, Wang L, Chen J, Nie D, Cong Y, Ahmad S, Ho A, Yuan P, Fung SH, Deng HH, et al. 2018. Craniomaxillofacial Bony Structures Segmentation from MRI with Deep-Supervision Adversarial Learning. In: Frangi AF, Schnabel JA, Davatzikos C, Alberola-López C, Fichtinger G, editors. Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. Cham: Springer International Publishing. (Lecture Notes in Computer Science). p. 720–727.

Liste des abréviations

- 2D : bidimensionnel
- 3D : tridimensionnel
- CBCT : *cone beam computed tomography* (tomographie volumique à faisceau conique)
- CI : *confidence interval* (intervalle de confiance)
- CMF : *craniomaxillofacial* (cranio-maxillo-facial)
- CNN : *convolutional neural network* (réseau neuronal convolutif)
- CT-Scan : *computed tomography scan* (tomodensitométrie ou TDM)
- DL : *deep learning* (apprentissage profond)
- DSC ou vDSC : *Dice Similarity Coefficient* (Indice de Sørensen-Dice)
- FH plane : *Francfort Horizontal plane* (plan horizontal de Francfort)
- FOV : *field of view* (champ d'acquisition)
- GPU : *Graphical Processing Unit* (carte graphique)
- IRM : imagerie à résonnance magnétique
- MRE : *mean radial error* (erreur radiale moyenne)
- MSP : *midsagittal plane* (plan sagittal median)
- R&R : *repeatability and reproducibility* (répétabilité et reproductibilité)
- ROI : *region of interest* (région d'intérêt)
- SCN : SpatialConfiguration-Net
- SD : *standard deviation* (écart type)
- SDR : *success detection rate* (taux de détection réussie)
- sDSC : *surface Dice Similarity Coefficient*
- stl : *STereoLithography*
- ULD : *ultra low-dose*

Table des figures

Table des tableaux

Publications et Communications

Publications

Dot G, Rafflenbeul F, Arbotto M, Gajny L, Rouch P, Schouman T. 2020. Accuracy and reliability of automatic three-dimensional cephalometric landmarking. International Journal of Oral and Maxillofacial Surgery. 49(10):1367–1378.

Dot G, Rafflenbeul F, Kerbrat A, Rouch P, Gajny L, Schouman T. 2021. Three-Dimensional Cephalometric Landmarking and Frankfort Horizontal Plane Construction: Reproducibility of Conventional and Novel Landmarks. Journal of Clinical Medicine. 10(22):5303.

Dot G, Schouman T, Dubois G, Rouch P, Gajny L. 2022. Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnU-Net framework. Eur Radiol. 32(6):3639–3648.

Dot G, Schouman T, Chang S, Rafflenbeul F, Kerbrat A, Rouch P, Gajny L. 2022 Aug 18. Automatic 3-Dimensional Cephalometric Landmarking via Deep Learning. J Dent Res.:00220345221112333.

Communications orales

Céphalométrie 3D automatisée: qu'en dit la littérature? Journées de l'Orthodontie, Paris Novembre 2019.

Automatic cephalometric landmarking of craniomaxillofacial computed tomography scans using a coarse-to-fine deep learning approach. IX International Conference on Computational Bioengineering, ICCB2022. Portugal, April 2022.

Reconstruction et annotation céphalométrique 3D de scanner pré-chirurgicaux : automatisation par apprentissage profond. 93e réunion scientifique de la Société Française d'Orthopédie Dento-Faciale, Lille Mai 2022 [communication orale invitée]

Segmentation et céphalométrie 3D automatisée par deep learning. Journées de l'Orthodontie, Paris Novembre 2022.

Communications affichées

Accuracy and reliability of automatic three-dimensional cephalometric landmarking: a systematic review. 96th Congress of the European Orthodontic Society, Hamburg 10-14 June 2020 [congrès annulé à cause de la situation sanitaire]

Fully automatic segmentation of computed tomography scans using deep learning: application to routine orthognathic surgery patients. 2021 Annual Virtual Conference of the European Orthodontic Society, 2-3 July 2021.

Automatic three-dimensional cephalometric landmarking of presurgical computed tomography scans via deep learning. 97th Congress of the European Orthodontic Society, Cyprus June 2022.

MATERIAUX

SUPPLEMENTAIRES

Matériaux Supplémentaires A : revue systématique de la littérature

Ces Matériaux Supplémentaires A comprend les matériaux supplémentaires de l'article présenté en Chapitre 2 : « Accuracy and reliability of automatic three-dimensional cephalometric landmarking. A systematic review ».

Supplementary Table A1. Search strategies used for various databases

Database	Search strategy	Results
MEDLINE via Pubmed	(tomography, x ray computed [mh] OR imaging, three-dimensional [mh] OR cone-beam* [tw] OR CBCT [tw] OR tomograph* [tw]) AND (cephalometry [mh] OR cephalometr* [tw] OR craniofacial [tw]) AND (Anatomic Landmarks [mh] OR landmark* [tw] OR point* [tw]) AND (algorithms [mh] OR Radiographic Image Interpretation, Computer-Assisted/methods [mh] OR algorithm* [tw] OR automat* [tw])	165
EMBASE	('three dimensional imaging'/exp OR 'computer assisted tomography'/exp OR cbct OR 'cone beam*' OR tomograph*) AND ('cephalometry'/exp OR cephalometr* OR craniometr* OR craniofacial) AND ('anatomic landmark'/exp OR landmark* OR point*) AND ('algorithm'/exp OR 'software'/exp OR algorithm*)	281
Web Of Science	Set #1: TS=(cone-beam computed tomography OR cone beam computed tomography OR CBCT OR tomograph* OR imaging, three-dimension OR 3D imaging) Set #2: TS=(cephalometr* OR craniometric* OR craniofacial) Set #3: TS=(landmark* OR point*) Set #4: TS=(algorithm* OR automat*) Set #5: #1 AND #2 AND #3 AND #4 Indexes=SCI-EXPANDED, SSCI, CPCI-S, CPCI-SSH, IC Timespan=All years	81
CENTRAL	([mh ^"tomography, x ray computed"] OR [mh ^"imaging, three-dimensional"] OR (cone-beam):kw OR CBCT:kw OR tomograph*:kw) AND ([mh cephalometry] OR cephalometr*:kw OR craniometric*:kw OR craniofacial:kw) AND ([mh ^"Anatomic Landmarks"] OR landmark*:kw OR point*:kw) AND ([mh algorithms] OR [mh ^"Radiographic Image Interpretation, Computer-Assisted"] OR algorithm*:kw OR automat*:kw)	1
OpenGrey	(cone-beam computed tomography OR cone beam computed tomography OR CBCT OR tomograph* OR imaging, three-dimension OR 3D imaging) AND (cephalometr* OR craniometric* OR craniofacial)	8
Google Scholar	3D AND cephalometric AND landmark AND automatic Sort by relevance, 300 first results	300

Supplementary Table A2. Tailored QUADAS-2 tool used for risk of bias and applicability evaluation

Domain 1: Patient selection	
A. Risk of bias	
Describe methods of patient selection:	
• Was a consecutive or random sample of patients enrolled?	Yes/No/Unclear
• Did the sample include an appropriate spectrum of objects (patients)?	Yes/No/Unclear
• Did the study avoid inappropriate exclusions?	Yes/No/Unclear
Could the selection of patients have introduced bias?	RISK: LOW/HIGH/UNCLEAR
B. Concerns regarding applicability	
Describe included patients (prior testing, presentation, intended use of index test and setting):	
Is there concern that the included patients do not match the review question?	CONCERN: LOW/HIGH/UNCLEAR
Domain 2: Index test	
A. Risk of bias	
Describe the index test and how it was conducted and interpreted:	
• Were the index test results interpreted without knowledge of the results of the reference standard?	Yes/No/Unclear
• Was algorithm described in sufficient detail to permit replication?	Yes/No/Unclear
• Was algorithm based on data independent from this patient selection?	Yes/No/Unclear
Could the conduct or interpretation of the index test have introduced bias?	RISK: LOW/HIGH/UNCLEAR
B. Concerns regarding applicability	
Is there concern that the index test, its conduct, or interpretation differ from the review question?	CONCERN: LOW/HIGH/UNCLEAR
Domain 3: Reference standard	
A. Risk of bias	
Describe the reference standard and how it was conducted and interpreted:	
• Is the reference standard likely to correctly classify the target condition (with description of observer reproducibility)?	Yes/No/Unclear

• Were the reference standard results interpreted without knowledge of the results of index test?	Yes/No/Unclear
• Was execution of reference standard described in sufficient detail to permit its replication?	Yes/No/Unclear
• Was observer reproducibility properly adressed and described?	Yes/No/Unclear
• For the in-vitro study, did the test situation imitate a clinical situation?	Yes/No/Unclear
Could the reference standard, its conduct, or its interpretation have introduced bias?	RISK: LOW/HIGH/UNCLEAR
B. Concerns regarding applicability	
Is there concern that the target condition as defined by the reference standard does not match the review question?	CONCERN: LOW/HIGH/UNCLEAR

Domain 4: Flow and timing

A. Risk of bias

Describe any patients who did not receive the index test(s) and/or reference standard:

• Did all patients receive a reference standard?	Yes/No/Unclear
• Did patients receive the same reference standard?	Yes/No/Unclear
• Were all patients included in the analysis?	Yes/No/Unclear
Could the patient flow have introduced bias?	RISK: LOW/HIGH/UNCLEAR

Supplementary Table A3. Mean localization error ± standard deviations (in mm) of reported landmarks

	Shahidi et al. 2014	Gupta et al. 2015	Zhang et al. 2016[a]	Codari et al. 2017[a]	Zhang et al. 2017[a,b]	De Jong et al. 2018	Montúfar et al. 2018	Neelapu et al. 2018	Torosdagli et al. 2018[a,c]	O'Neil et al. 2019[a,d]
Point A	3.11 ± 0.74	1.73 ± 0.80		1,80 ± 0,86			1.46 ± 0.75	1.91 ± 0.94		
Point B	3.86 ± 1.41	2.08 ± 1.09		2,66 ± 1,33			2.53 ± 0.56	1.78 ± 0.91	0.34 ± 0.72	
Anterior Nasal Spine (ANS)	3.12 ± 0.80	1.42 ± 0.73		2.58 ± 1.50		5.6 ± 8.1	1.72 ± 0.91	1.03 ± 0.62		Observer A: 2.57 ± 3.37 Observer B 2.84 ± 3.49
Posterior Nasal Spine (PNS)	3.60 ± 1.35	2.08 ± 1.29		1.64 ± 1.18			2.17 ± 1.27	1.60 ± 1.15		
Pogonion (Pog)	3.00 ± 1.02	1.53 ± 0.79	1.03 ± 0.53	2.88 ± 1.52	0.93 ± 0.47	4.6 ± 8.4	2.59 ± 0.98	1.77 ± 0.96	1.55 ± 1.98	
Nasion (N)	3.20 ± 1.64	1.17 ± 0.49	1.62 ± 0.82	3.19 ± 3.33	0.96 ± 0.69	3.0 ± 2.5	2.14 ± 1.04	0.95 ± 0.69		Observer A: 2.35 ± 1.48 Observer B: 4.04 ± 2.10
Sella (S)	3.45 ± 1.82	1.52 ± 0.75		1.44 ± 0.73			2.67 ± 2.05	2.19 ± 0.91		
Gnathion (Gn)	3.77 ± 2.69	1.62 ± 0.62					2.10 ± 1.06	1.64 ± 0.68	0.49 ± 1.42	
Menton (Me)	3.59 ± 1.79	1.21 ± 0.58	1.02 ± 0.73	1.76 ± 0.83	0.81 ± 0.71		2.28 ± 1.15	1.57 ± 0.54	0.04 ± 0.12	
Gonion Left (GoL)		2.04 ± 1.47	1.59 ± 0.88	3.92 ± 2.38	1.51 ± 1.00	2.6 ± 2.0	2.33 ± 1.62	2.02 ± 1.09		
Gonion Right (GoR)		2.47 ± 1.37	1.61 ± 1.11	3.20 ± 1.96	1.79 ± 0.65	4.8 ± 5.7	2.45 ± 1.76	2.10 ± 1.18		
Condylar Left (CdL)		3.20 ± 2.49						3.78 ± 2.77	0.34 ± 0.60	
Condylar Right (CdR)		2.38 ± 1.71						3.34 ± 2.47	0.08 ± 0.24	
Frontozygomatic Left (FzL)		1.47 ± 0.86		2.84 ± 2.36		2.0 ± 1.2				
Frontozygomatic Right (FzR)		1.60 ± 0.71		2.54 ± 1.76		1.5 ± 1.1				
Lateral Zygomatic Left (LatzL)		2.80 ± 1.63				2.1 ± 1.1		1.74 ± 1.01		
Lateral Zygomatic Right (LatzR)		2.83 ± 2.05				1.7 ± 1.0		1.48 ± 1.05		
Orbitale Left (OrL)		1.78 ± 1.36	1.55 ± 0.70	1.74 ± 1.08	1.08 ± 0.53	1.9 ± 2.5	3.12 ± 2.70			
Orbitale Right (OrR)		2.37 ± 2.23	1.58 ± 0.85	1.69 ± 1.28	0.97 ± 0.56	3.7 ± 3.4	3.46 ± 2.13			
Basion (Ba)				2.13 ± 0.62			1.78 ± 1.62			
Gonion (Go)	3.72 ± 1.67									
Upper Incisive Tip (U1T)	3.59 ± 1.76						1.88 ± 0.81			
Upper Incisive Tip Left (U1TL)			1.30 ± 0.90	1.62 ± 0.91	1.32 ± 0.70					
Upper Incisive Tip Right (U1TR)			1.10 ± 0.65	2.23 ± 1.41	1.01 ± 0.81					
Upper Incisive Apex (U1A)	3.15 ± 0.91									
Lower Incisive Tip (L1T)	3.30 ± 0.92						3.19 ± 0.81			
Lower Incisive Tip Left (L1TL)			1.49 ± 0.61	2.26 ± 0.89	1.20 ± 1.28					
Lower Incisive Tip Right (L1TR)			1.40 ± 0.55	3.52 ± 2.22	0.98 ± 0.56					
Lower Incisive Apex (L1A)	3.08 ± 1.08									
Infradentale (Id)						3.7 ± 3.0			0.52 ± 0.96	
Supradentale (Sd)						3.7 ± 4.0				
Porion Left (PoL)							3.64 ± 3.93			
Porion Right (PoR)							3.72 ± 3.76			
Coronoid Left (CoL)						5.2 ± 7.0			0 ± 0	
Coronoid Right (CoR)						8.0 ± 7.9			0.45 ± 1.43	
Mental foramen Left						1.4 ± 1.8				

Mental foramen Right			1.0 ± 0.8	
Supraorbital foramen Left			2.5 ± 2.1	
Supraorbital foramen Right			2.9 ± 2.4	
Transition nasal-frontal			2.7 ± 1.6	
Anterior nasal			1.7 ± 1.4	
Lateral nasal aperture Right			5.6 ± 9.8	
Lateral nasal aperture Left			2.8 ± 4.1	
Distal nasal aperture Right			1.2 ± 0.8	
Distal nasal aperture Left			2.2 ± 4.4	
Maxilla-zygomatic transition Left			4.9 ± 3.8	
Maxilla-zygomatic transition Right			4.5 ± 3.1	
Corner zygomatic Right			1.2 ± 0.6	
Corner zygomatic Left			1.1 ± 0.7	
Zygomatic process Right			2.6 ± 2.5	
Zygomatic process Left			14.4 ± 31.1	
Distal mandibular notch Left		2.05 ± 1.22	13.7 ± 29	
Distal mandibular notch Right		2.26 ± 1.07	8.8 ± 7	
Jugal Left		1.70 ± 0.71		
Jugal Right		1.59 ± 0.65		
Superiorposterior extremity C2 (C2sp)				1.59 ± 0.92
Anteriorinferior point C3 (C3ai)				1.70 ± 1.05
Anteriorinferior point C4 (C4ai)				1.95 ± 1.25
Posterior maxillary point Left (PML)		1.85 ± 0.96		
Posterior maxillary point Right (PMR)		2.53 ± 1.85		
Deepest point anterior border of ramus Left (R1L)				1.40 ± 0.81
Deepest point anterior border of ramus Right (R1R)				2.10 ± 1.12
Lower second molar Left	1.24 ± 0.62	0.86 ± 0.91		
Lower second molar Right	1.63 ± 0.75	0.96 ± 0.44		
Upper second molar Left	1.81 ± 1.03	0.97 ± 0.61		
Upper second molar Right	1.59 ± 0.80	1.13 ± 0.68		

[a] Unpublished data shared by the authors [b] Results for “JSD” method

[c] Results for “max pool without dropout” [d] Results for Observer A

Matériaux Supplémentaires B : segmentation automatisée

Ces Matériaux Supplémentaires B comprend les matériaux supplémentaires de l'article présenté en Chapitre 4 : « Fully automatic segmentation of craniomaxillofacial CT scans for computer-assisted orthognathic surgery planning using the nnU-Net framework ».

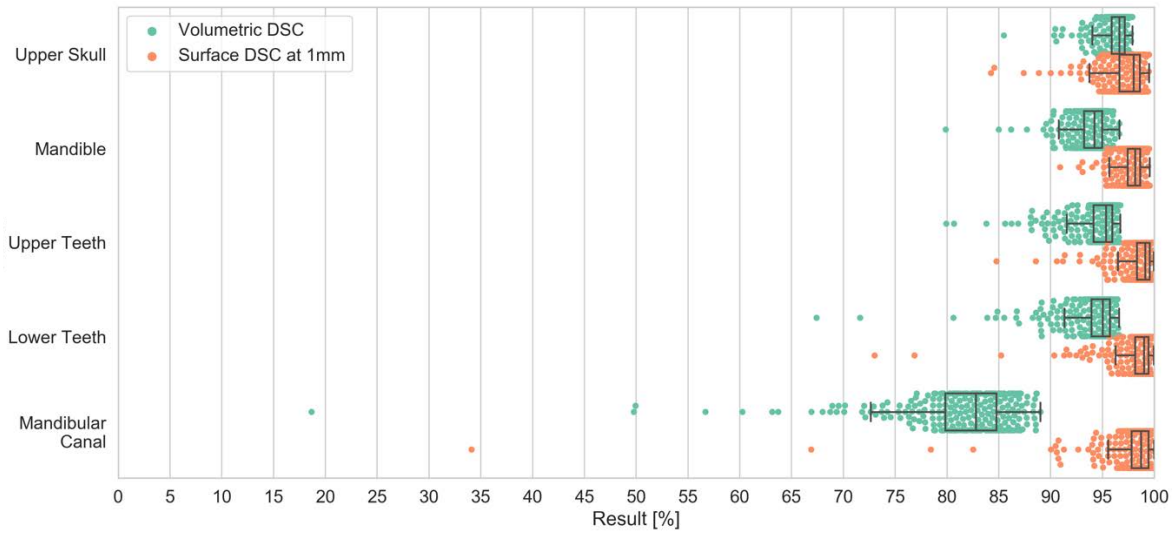
Implementation Details.

All experiments were performed using the nnU-Net v.1.6.5 framework running on an Nvidia Pytorch Docker container (v.20.10-py3) on our laboratory workstation (CPU AMD Ryzen 9 3900X 12-Core; 128 Gb RAM; GPU Nvidia Titan RTX 24Gb). Our preliminary tests showed that the 3D U-Net full-resolution model far outperformed the 2D U-Net model, while 3D U-Net cascade performances were inconsistent. As a result, we focused on training the 3D U-Net full-resolution model. Target spacing of the model was 0.31*0.45*0.45mm, patch size was 192*112*112 pixels and batch size was set to 2.

Training was performed once on our train/validation set following a 5-fold cross-validation strategy. Training time for one fold was about 48 hours (1,000 epochs) and the GPU VRAM memory footprint was about 8GB. Our automatic post-processing strategy showed that removing all but the largest components improved performance in segmentation masks of upper skull and mandible. After the end of the training pipeline, we assessed the need for additional post-processing by looking at the predictions with the worst results. Analysis of cross-validation results showed that the model incorrectly labeled a few voxels as teeth in some scans displaying no upper and/or lower teeth. As a result, we implemented further post-processing for teeth masks, using SimpleITK library to remove all components smaller than a threshold which we empirically set at 60mm^3, i.e. the largest volume that improved cross-validation results. Cross-validation quantitative results (300 CT scans) are provided in Supplementary Table 1 and Supplementary Figure 1. Inference was performed once with TTA, then once without, on the test datasets. Inference took approximately 45 minutes per CT scan with TTA, compared to 10 minutes without. Since disabling TTA did not negatively affect the prediction results, we chose to present results for inference without TTA. The post-processing strategy described above was applied to the prediction results.

Our quantitative metrics were computed using SimpleITK library v.2.0.2 (for vDSC, Jaccard Coefficient, Volumetric Similarity, Average Surface Distance and Hausdorff distance) and Medical Segmentation Decathlon Challenge implementation (for sDSC).

For the public mandible dataset, an additional fully-automatic cavity filling was performed using the 3D Slicer software "wrap solidify" filter (outer surface extraction, minimum set to 10mm) and mandible mask predictions were analyzed.



Supplementary Figure B1. Volumetric DSC and surface DSC at 1mm *results of our* 5-fold cross-validation *(300 CT scans).*

Supplementary Table B1. Mean ± Standard Deviation quantitative results of our 5-fold cross-validation (300 CT scans)

	Upper Skull	Mandible	Upper teeth	Lower Teeth	Mandibular canal	Total
Volume DSC	0.9632 ± 0.0138	0.9386 ± 0.0179	0.9455 ± 0.024	0.9414 ± 0.0309	0.814 ± 0.0656	0.9204 ± 0.0648
Jaccard Coefficient	0.9294 ± 0.0248	0.8848 ± 0.0305	0.8976 ± 0.0407	0.8907 ± 0.0495	0.6906 ± 0.0788	0.6773 ± 0.3595
Volume Similarity	-0.0043 ± 0.0421	-0.0136 ± 0.0612	-0.0194 ± 0.0583	-0.0188 ± 0.0683	-0.0160 ± 0.0799	0.1668 ± 0.3678
Average Surface Distance (GT to Prediction)	0.1491 ± 0.098	0.1167 ± 0.0546	0.1021 ± 0.0825	0.1049 ± 0.0985	0.1985 ± 0.5738	0.1344 ± 0.2703
Average Surface Distance (Prediction to GT)	0.0698 ± 0.0677	0.0866 ± 0.0556	0.0909 ± 0.2332	0.1161 ± 0.4811	0.1646 ± 0.2508	0.1056 ± 0.2684
Hausdorff Distance 100% (mm)	9.6152 ± 3.9791	4.0386 ± 1.7017	3.9237 ± 7.6032	3.8274 ± 7.1537	3.341 ± 2.7371	4.9541 ± 5.6882
Hausdorff Distance 95% (mm)	0.8863 ± 0.7018	0.6942 ± 0.2767	0.5909 ± 0.3977	0.8066 ± 2.7852	0.8845 ± 1.9401	0.7731 ± 1.5674
Surface DSC at 1mm	0.9736 ± 0.0205	0.9791 ± 0.0117	0.9860 ± 0.0176	0.9832 ± 0.0262	0.9786 ± 0.0473	0.9801 ± 0.0278

Supplementary Table B2. Mean ± Standard Deviation quantitative results on our test dataset (153 CT scans)

	Upper Skull	Mandible	Upper teeth	Lower Teeth	Mandibular canal	Total
Volume DSC	0.9622 ± 0.0143	0.9419 ± 0.0162	0.9483 ± 0.0181	0.9438 ± 0.0232	0.8159 ± 0.0579	0.9224 ± 0.0619
Jaccard Coefficient	0.9274 ± 0.0258	0.8907 ± 0.0279	0.9022 ± 0.0316	0.8944 ± 0.0389	0.6928 ± 0.076	0.8615 ± 0.0961
Volume Similarity	-0.0133 ± 0.0412	-0.0118 ± 0.0571	-0.0183 ± 0.0527	-0.0177 ± 0.0606	-0.0119 ± 0.0819	-0.0146 ± 0.0601
Average Surface Distance (GT to Prediction)	0.1695 ± 0.138	0.1137 ± 0.0508	0.0912 ± 0.0537	0.102 ± 0.1119	0.1935 ± 0.2908	0.134 ± 0.1608
Average Surface Distance (Prediction to GT)	0.0625 ± 0.0421	0.0807 ± 0.0488	0.0972 ± 0.3492	0.0779 ± 0.0697	0.157 ± 0.1364	0.0951 ± 0.1754
Hausdorff Distance 100% (mm)	9.4477 ± 4.2425	4.0158 ± 1.5094	4.6572 ± 18.8037	3.5852 ± 2.7675	3.4583 ± 2.933	5.0328 ± 9.078
Hausdorff Distance 95% (mm)	1.0097 ± 0.8568	0.697 ± 0.2868	0.5403 ± 0.2273	0.589 ± 0.5276	0.9853 ± 1.9992	0.7642 ± 1.0317
Surface DSC at 1mm	0.9692 ± 0.0308	0.9792 ± 0.0122	0.9887 ± 0.0118	0.9853 ± 0.02	0.979 ± 0.0351	0.9803 ± 0.0248

Supplementary Table B3. Quantitative results on public mandible test dataset (10 CT scans)

	Prediction vs Operator	Mandible number									
		1	2	3	4	5	6	7	8	9	10
Volume DSC	A	0.9172	0.8512	0.8929	0.8938	0.9216	0.9074	0.9215	0.8871	0.9266	0.6119
	B	0.9147	0.8468	0.8980	0.8911	0.9203	0.9129	0.9277	0.8823	0.9257	0.6146
Jaccard Coefficient	A	0.8471	0.7409	0.8065	0.8079	0.8545	0.8306	0.8544	0.7971	0.8632	0.4408
	B	0.8428	0.7344	0.8149	0.8037	0.8524	0.8398	0.8652	0.7894	0.8617	0.4437
Volume Similarity	A	0.0410	0.0200	0.0325	0.0175	0.0200	-0.0135	0.0070	0.1420	-0.0234	0.6886
	B	0.0041	0.0411	0.0720	0.0370	0.0185	-0.0013	-0.0072	0.1641	-0.0315	0.6802
Average Surface Distance (GT to Prediction)	A	0.2620	0.4735	0.2562	0.3892	0.2329	0.2820	0.2834	0.4199	0.2221	26.8738
	B	0.2644	0.4933	0.2382	0.4205	0.2385	0.2593	0.2573	0.4463	0.2256	27.0208
Average Surface Distance (Prediction to GT)	A	0.2201	0.2138	0.1908	0.1846	0.1721	0.2256	0.1557	0.2437	0.1789	0.2053
	B	0.2302	0.2277	0.1717	0.1972	0.1753	0.2176	0.1687	0.2591	0.1775	0.2063
Hausdorff Distance 100% (mm)	A	4.0352	11.3262	4.0872	5.1444	2.5488	2.8038	3.8194	6.9552	2.9770	90.3069
	B	4.0000	10.9110	3.1053	4.2960	2.5993	3.0000	3.7950	6.9552	2.9770	90.4536
Hausdorff Distance 95% (mm)	A	1.0000	2.0000	1.2598	1.9994	0.7209	1.3037	1.0000	2.0000	0.7500	80.9020
	B	1.0000	2.0615	0.9390	1.9994	1.0000	1.0307	1.0000	2.0000	0.7500	81.0128
Surface DSC at 1mm	A	0.9710	0.9296	0.9562	0.9056	0.9803	0.9412	0.9729	0.9084	0.9910	0.6386
	B	0.9708	0.9232	0.9636	0.8952	0.9775	0.9481	0.9801	0.9002	0.9903	0.6404

Matériaux Supplémentaires C : reproductibilité du placement manuel des points céphalométriques

Ces Matériaux Supplémentaires C comprend les matériaux supplémentaires de l'article présenté en Chapitre 6 : « Three-Dimensional Cephalometric Landmarking and Frankfort Horizontal Plane Construction: Reproducibility of Conventional and Novel Landmarks ».

Supplementary Material C1 : Written instructions for landmarks process. Mimics software

Useful Mimics commands:

3D view	Rotation of the object	Mouse right click + move near the object
	Tilt of the object	Mouse right click + move far from the object
	Translation of the object	Shift + right mouse click + move
	Zoom in/out	Ctrl + right mouse click + move or Mouse scroll
2D views	Translation of the slice	Shift + right mouse click + move
	Zoom in/out	Ctrl + right mouse click + move
	Change slice	Mouse scroll
	Change contrast	Alt + right mouse click + move

2

Procedure

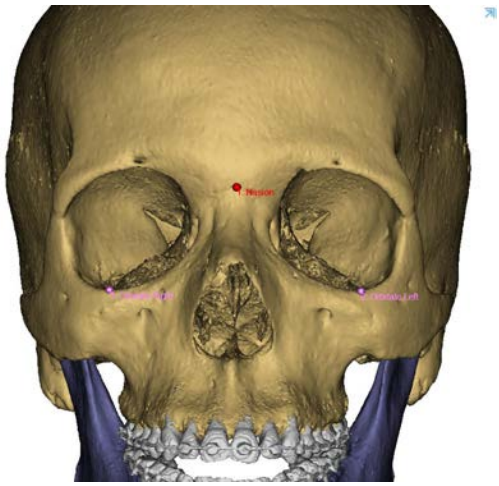
- Open Mimics document
- If needed, change document's layout to get a bigger 3D view: View > Layouts > Horizontal
- On the right panel, hide all masks (eye icon) & all objects other than Upper Skull / Mandibula / Upper Teeth / Lower Teeth
- Orient the skull object for it to be approximately aligned with Frankfort plane
- Launch Analyze > Measure and Analyze and choose "Full ceph analysis"

- Start Time measurement (chronometer)

- Follow the order of the landmarks to annotate (1:, 2:, 3:.... to 33:)
- If a landmark is missing (e.g. missing tooth), do not annotate it

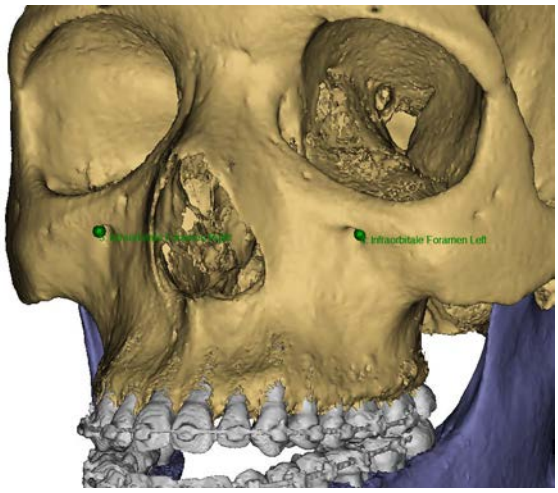
3

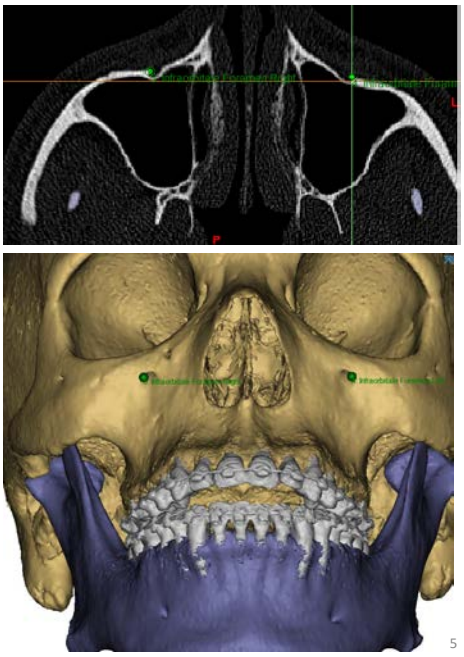
1: Nasion	Medial (and upper) point of the fronto-nasal suture	3D view, frontal Check on 2D slices if necessary (axial/sagittal)
2,3: Orbitale L/R	Lowest point of the orbitale rim L/R	3D view, frontal and check from above that the points are well on the orbital rim



4

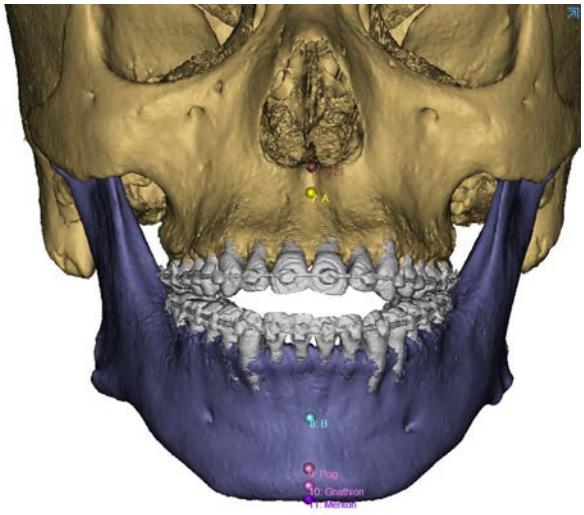
4,5: Infraorbital Foramen L/R	External & most distal point of the infraorbital foramen L/R	3D view, ¾ mesio-lateral view Check on 2D slice (Axial)





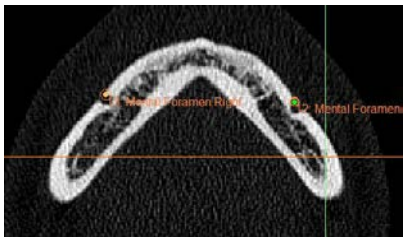
5

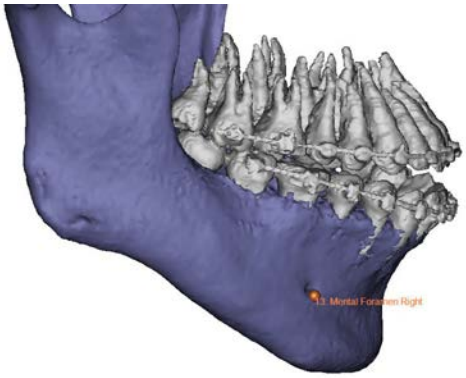
6: ANS	Medial and most anterior point of the nasal spine	3D view, frontal Check on 2D slices (axial/sagittal) In case of bifid spine, stay at a medial position
7: A	Medial & most posterior point of the maxilla	3D view, frontal Check afterwards laterally (see after point 18)
8: B	Medial & most osterior point of the mandible	3D view, frontal Check afterwards laterally (see after point 18)
9: Pog	Medial and most anterior point of the mandible	3D view, frontal Check afterwards laterally (see after point 18)
10: Gnathion	Medial & mid-point between Pog and Me	3D view, frontal Check afterwards laterally (see after point 18)
11: Menton	Medial and lowest point of the mandible	3D view, inferior Check afterwards laterally (see after point 18)

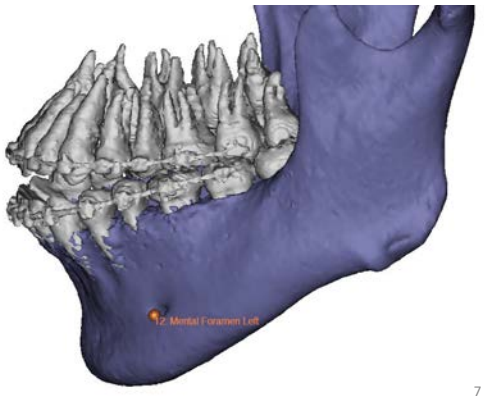


6

12, 13: Mental Foramen L/R	External most mesial point of the mental foramen L/R	3D view, ¾ disto-lateral view Check on 2D slice (Axial)

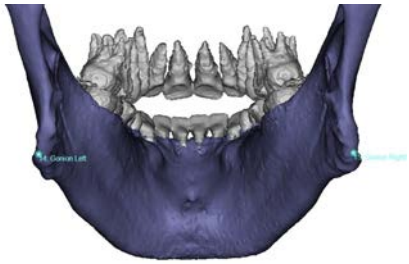


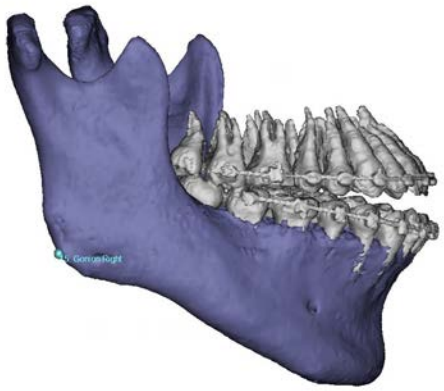


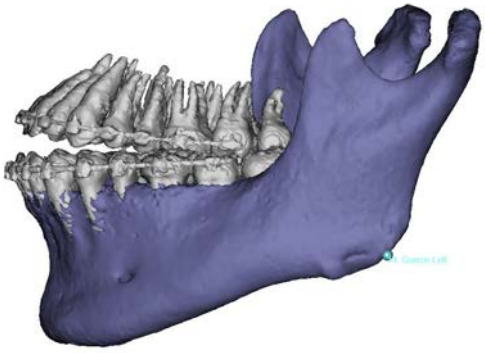


7

14,15: Gonion L/R	Mid-point of the gonial angle L/R	3D view, lateral. Mid-point between the beginning and the end of the mandibular angle Check from behind that the points are well located on the mandibular external border Before placing R one it can be easier to hide L one

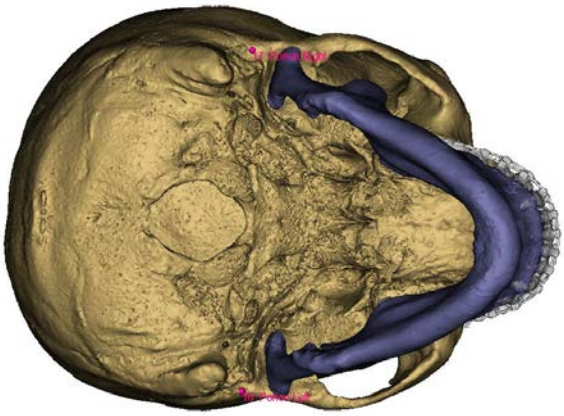


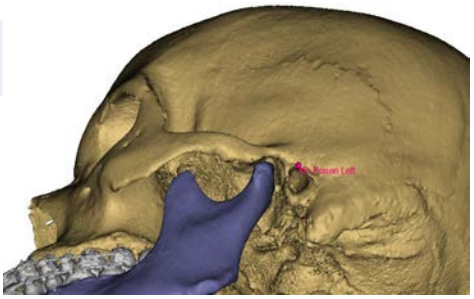


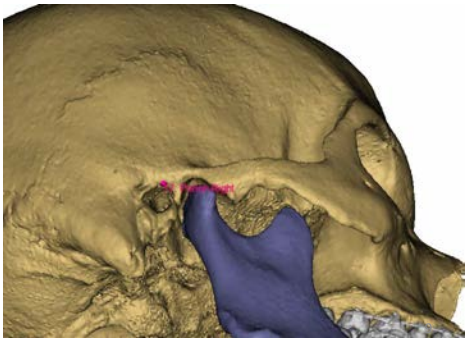


8

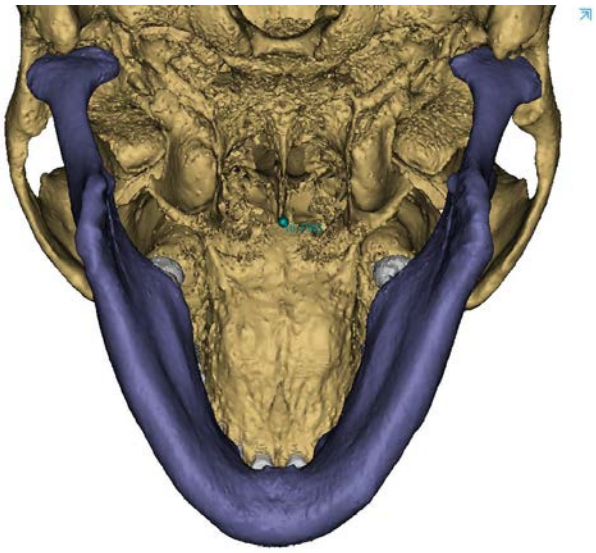
16, 17: Porion L/R	External & uppermost point of the auditory canal L/R	3D view, inferior. At the entry of auditory canal

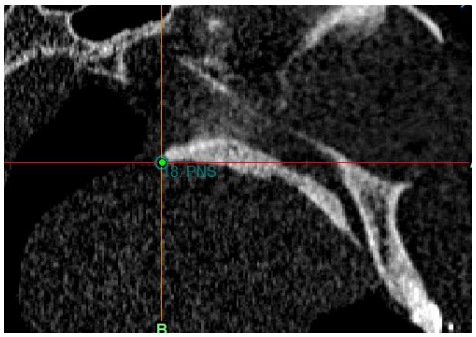






18: PNS	Medial & most distal point of the osseous palate	3D view, inferior-posterior Check on 2D slice (sagittal)





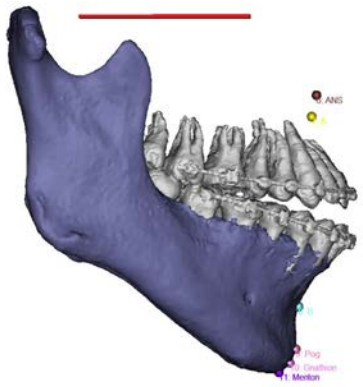
10

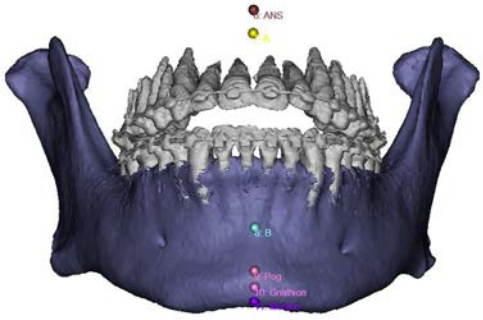
In the right lateral panel, hide the skull and align the 3D lateral view on the Frankfort plane automatically shown in red. **Check the landmarks 7, 8, 9, 10, 11 laterally:**

1/ 3D view, lateral, reorient to have Frankfort plane horizontal
For A point, show the skull
Check position of the points, correct them if necessary.
This can be done on the 2D slices (sagittal)

2/ 3D view, frontal, check that the medial position and alignment of the points is still OK

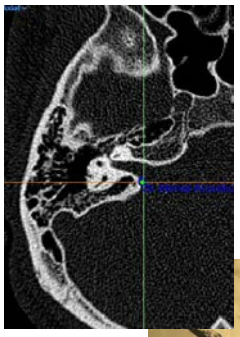
Afterwards, hide Frankfort plane (glasses icon) and show the skull

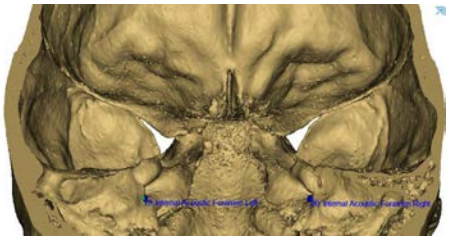


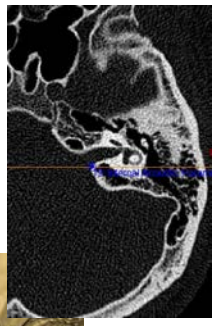


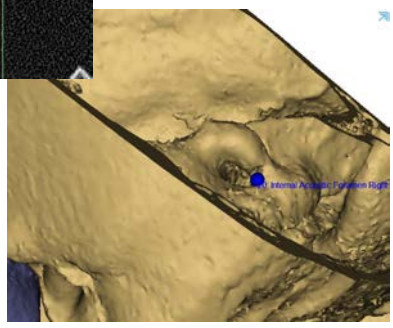
11

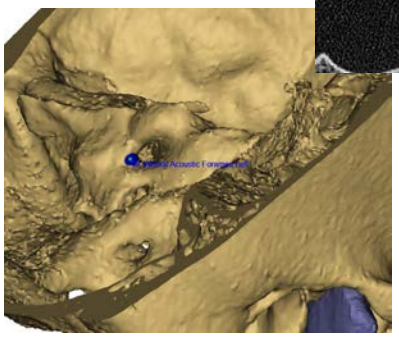
19, 20: Internal Acoustic Foramen L/R	External, most mesial and posterior point of the acoustic foramen L/R	3D view, ¾ postero-mesio-lateral view Check on 2D slice (Axial)





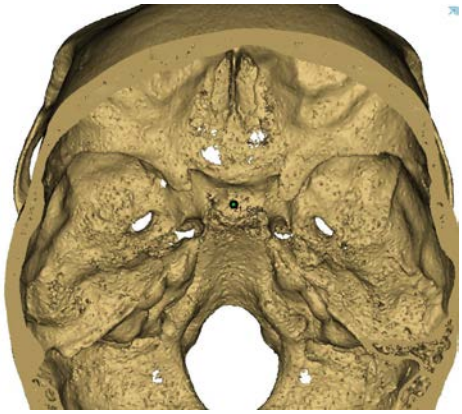


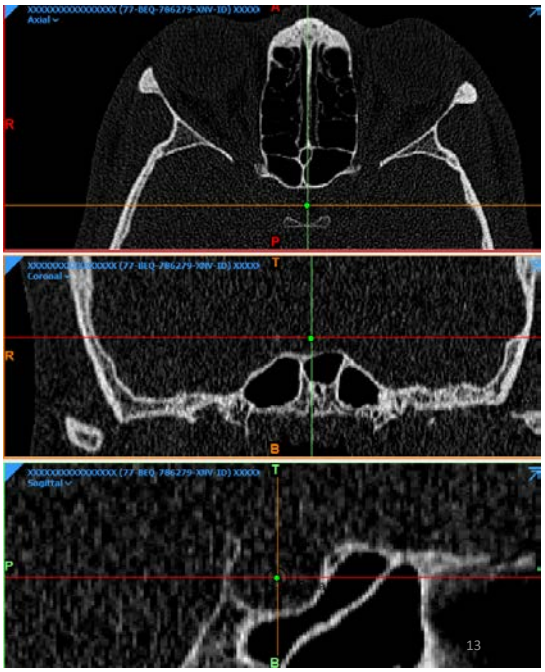




12

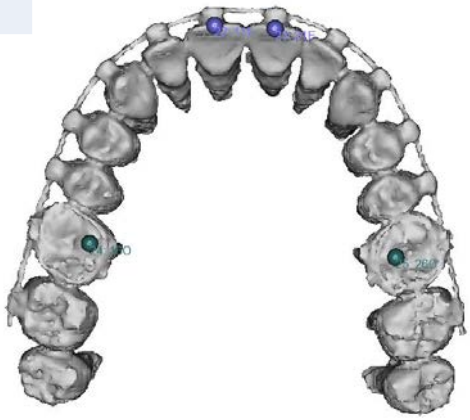
21: Sella	Central point of the sella	2D slices, first sagittal and refine on axial & coronal slices





Hide all the landmarks localized (glasses icon)
Hide Upper Skull, Mandible and Lower teeth objects

22, 23: 11E, 21E	Mid-point of 11/21 incisal edges	3D view, inferior
24,25: 16O, 26O	Summit of the mesio-palatal cusp 16/26	3D view, inferior



26, 27: 11A, 21A	Root apex of 11/21	3D view, superior

15

28, 29: 31E, 41E	Mid-point of 31/41 incisal edgess	3D view, superior
30, 31: 360, 460	Central fossa of 36/46	3D view, superior

16

32, 33: 31A, 41A	Root apex of 31/41	3D view, inferior

17

- Stop Time measurement (chronometer)
- Export the xml file using the “export” option of Measure & Analyze window

Thank you ☺

18

Supplementary Materials C2: Analysis of outliers

Method. For each landmark, repeatability and reproducibility standard deviations (SD) were computed according to the ISO 5725 standard (ISO 5725-2:2019) of the International Organization for Standardization. Upon first inspection of the results, the standard’s recommendations were followed for clear outlier points, whose annotations were considered as missing data.

Results. We inspected all modified Bland-Altman graphs in order to find possible clear outliers. The only outliers were found for mental foramen points right/left localized by operator #3 during the first annotation session (subjects 4 to 20). Figure SM1 shows Bland-Altmann graphs of the results with these outliers.

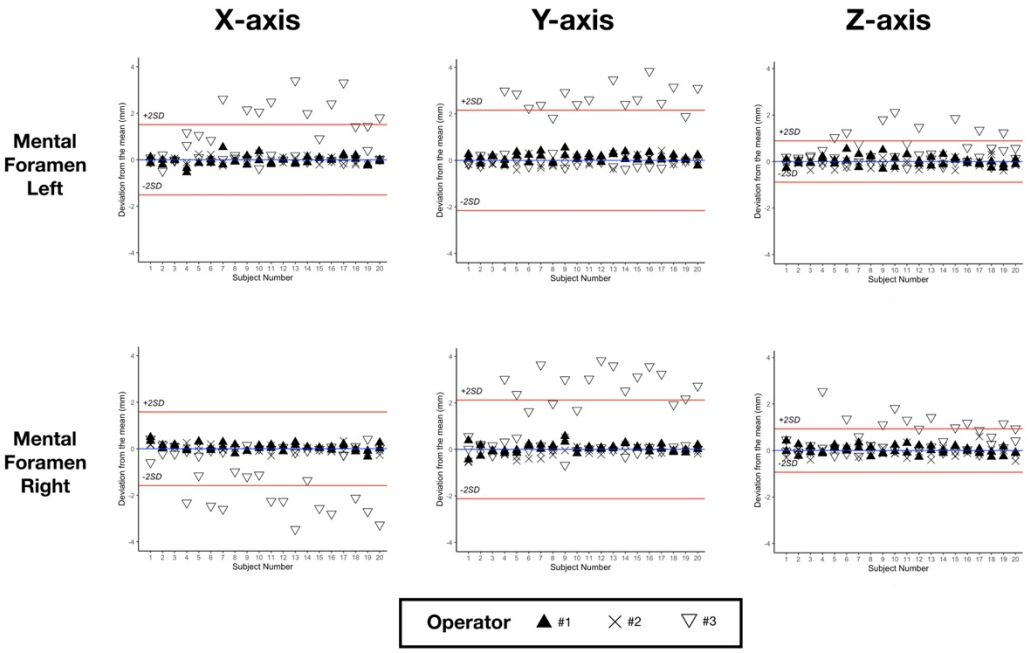


Figure SM1: Bland-Altman plots for mental foramen left and right with outliers, showing the deviations from the mean (blue line) of the 6 repetitions for the 20 subjects. Red lines show the ± 2*SD of reproducibility. SD, standard deviation.

Qualitative analysis of the results showed that the concerned landmarks had been localized in the most distal part of the mental foramina, contrary to the agreed upon definition which demanded placing these landmarks on the most mesial side of the foramen. Consistently with ISO 5725 recommendations, the corresponding annotations were treated as missing data. Figure SM2 shows Bland-Altmann graphs of the results after removal of the outliers.

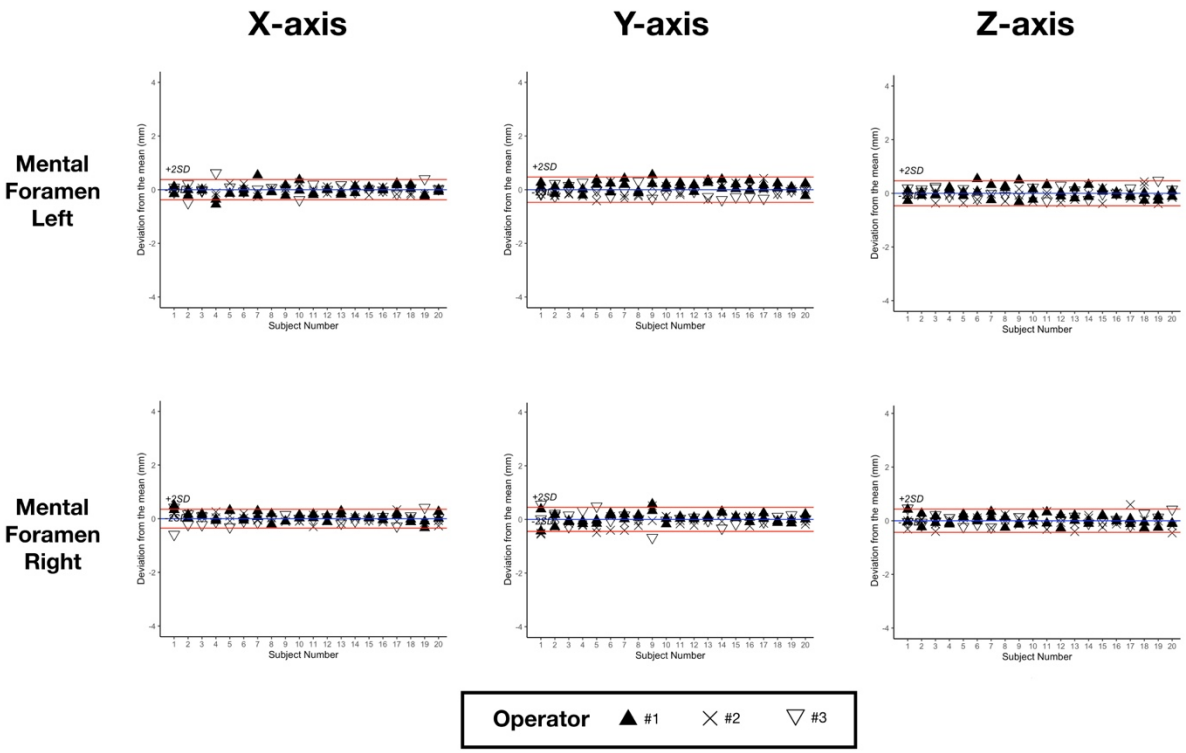


Figure SM2: Bland-Altman plots for mental foramen left and right without outliers, showing the deviations from the mean (blue line) of the 6 repetitions for the 20 subjects. Red lines show the ± 2*SD of reproducibility. SD, standard deviation.

Supplementary Materials C3: Angular distances between conventional and novel FH planes

Method. For each CT scan, four planes were computed using the means of all our operators' observations. Plane definition and labelling followed Pittayapat *et al.*'s publication: FH 1, FH 2, Plane 1 and Plane 3 (Table ST1). The absolute angular differences between each pair of planes were then computed, using trigonometry to calculate the angles between the normals to the planes.

Supplementary Table C1: Definition of the 4 planes used in supplementary analysis

Plane	**Definition**
Frankfort horizontal plane 1 (FH 1)	FH by connecting mid-Po, Or-R and Or-L
Frankfort horizontal plane 2 (FH 2)	FH by connecting mid-Or, Po-R and Po-L
Plane 1	A plane connecting mid-Or, IAF-R, IAF-L
Plane 3	A plane connecting Or-R, Or-L and mid-IAF

Results. Absolute angular distances between each pair of planes are summarized in Table ST2.

Supplementary Table C2: Mean absolute angular differences and standard deviations between each pair of planes. FH, Frankfort horizontal plane; SD, standard deviation

	Angular measurement (°)	
Plane	Mean	SD
FH1 – FH2	0.98	0.57
FH1 – Plane 1	2.94	1.34
FH1 – Plane 3	2.04	1.34
FH2 – Plane 1	2.59	1.27
FH2 – Plane 3	2.41	1.21

Matériaux Supplémentaires D : placement automatisé des points céphalométriques

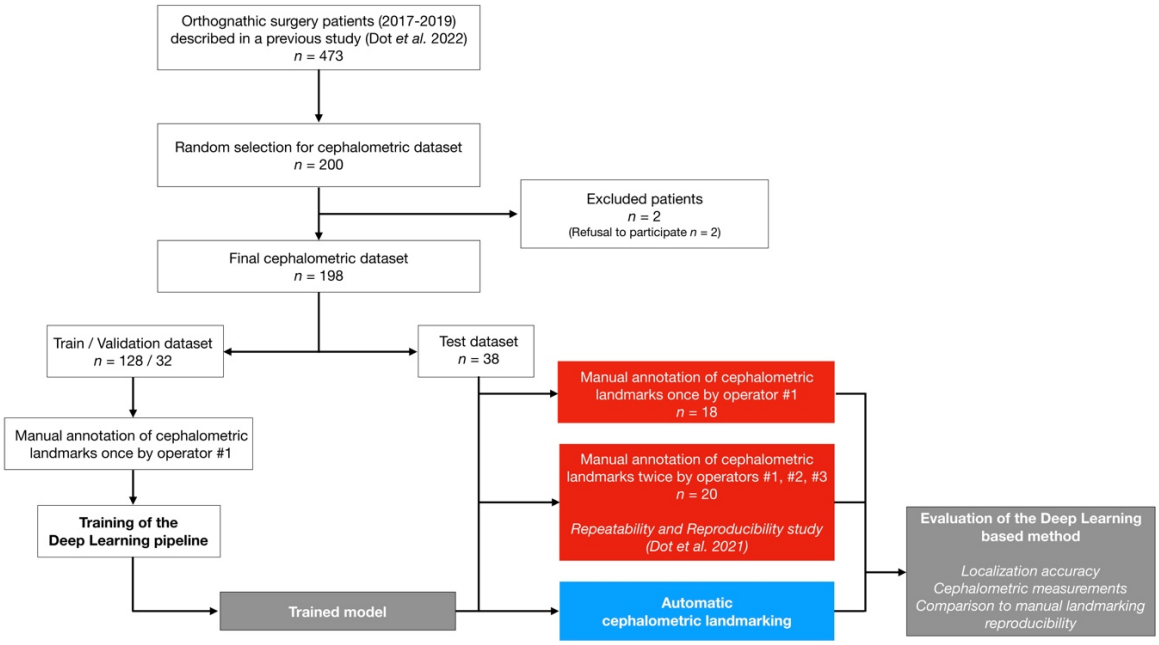
Ces Matériaux Supplémentaires D comprend les matériaux supplémentaires de l'article présenté en Chapitre 7 : « Three-Dimensional Cephalometric Landmarking and Analysis of Craniomaxillofacial CT scans via Deep Learning».

Materials and Methods

A DL pipeline was followed to localize cephalometric landmarks automatically on randomly selected craniomaxillofacial CT scans. We compared the results of our DL-based method on a hold-out test dataset (the index test) with those obtained by manual landmarking (the reference test). Our outcome set included localization accuracy, cephalometric measurements and comparison to manual landmarking reproducibility. The data were analyzed using R (http://www.r-project.org) and Python (http://www.python.org, version 3.7).

Dataset

Data were selected from a retrospective cohort of all consecutive patients having undergone orthognathic surgery in a single maxillofacial surgery department between January 2017 and December 2019, as described in a previous study (Dot et al. 2022). Patients referred to this center presented a wide variety of dentofacial deformities, came from various socioeconomic backgrounds, and were ethnically diverse. Patients were considered for inclusion whatever dental deformity they presented, with no minimum age. Exclusion criteria were refusal to participate in the research (all patients were contacted by mail). 200 subjects were randomly selected from this cohort, and 2 patients refused to participate. 198 subjects were eventually included in our dataset. The 198 CT scans performed before included subjects' orthognathic surgeries were de-identified and given an anonymization code (the anonymization chart was kept by the clinical investigator), no personal data was entered into the algorithm.



Supplementary Figure D1. Data flow of patient selection, training and evaluation process. Red, reference test; blue, index test.

The vast majority of the included CT scans (*n* = 185, 93.4%) were acquired on a Discovery CT750 HD scanner (GE Healthcare, Chicago, USA) set at 100kVp, 50mAs, exposure time 730ms, slice thickness 0.625mm and slice increment 0.320mm. Field of view ranged from 165 to 320mm and pixel size ranged from 0.32 to 0.63mm. CT scans and patient characteristics are detailed in Supplementary Table D1.

Supplementary Table D1. CT scans and patient characteristics for the train/validation/test dataset.

	Train	**Validation**	**Test**
Number of CT scans	128	32	38
Age, mean ± SD, years	28 ± 11	28 ± 11	26 ± 8
Gender, no. (%)			
Female	76 (59.3)	20 (62.5)	20 (52.6)
Male	52 (40.7)	12 (37.5)	18 (47.4)
Skeletal deformity, no. (%)			
Class I	15 (11.7)	3 (9.3)	3 (7.9)
Class II[a]	65 (50.8)	17 (53.1)	22 (57.9)
Class III[b]	48 (37.5)	12 (37.5)	13 (34.2)
Syndromic deformity	6 (4.7)	2 (6.3)	3 (7.9)
Metal artifacts, no. (%)			
Orthodontic materials	104 (81.3)	23 (71.9)	33 (86.8)
Metallic dental filling/crown	55 (43.0)	14 (43.8)	16 (42.1)
No metallic artifact	14 (10.9)	4 (12.5)	3 (7.9)
Mean in-plane pixel size (mm^2)	0.44 * 0.44	0.45 * 0.45	0.46 * 0.46
Mean field of view (mm)	228	229	233
Mean slice thickness (mm)	0.33	0.33	0.32

Mean number of slices	742	732	762
Number of scans by CT Machine			
GEHC Discovery CT750 HD	121	32	36
GEHC Optima CT660	3		1
Other CT Machine[c]	4		1

[a]Prognathic maxilla and/or retrognathic mandible. [b]Retrognathic maxilla and/or prognathic mandible. [c]GEHC Revolution, Philips Ingenuity, Siemens SOMATOM. SD, Standard Deviation; GEHC: GE Healthcare.

Manual Landmarking (Reference Test)

Thirty-three commonly used landmarks, divided into skeletal (*n* = 21) and dental (*n* = 12), were manually annotated on each CT scan on the software Mimics (v.22.0, Materialise, Leuven, Belgium), following written and verbal instructions on the 3D description and annotation procedure for each landmark (please refer to (Dot et al. 2021) for more details about the written instructions provided to the operators). Definitions of the skeletal and dental landmarks are provided in Supplementary Table D2 and D3, respectively.

Supplementary Table D2. Definition of the "skeletal" landmarks localized in our study (L/R: Left/Right)

Landmark name	Description
Nasion (Na)	Medial (and upper) point of the frontonasal suture
Orbitale L/R (Or-L / Or-R)	Lowest point of the orbital rim L/R
Anterior Nasal Spine (ANS)	Medial and most anterior point of the nasal spine
A Point (A)	Medial and most posterior point of the anterior concavity of the maxilla
B Point (B)	Medial and most posterior point of the anterior concavity of the mandible
Pogonion (Pog)	Medial and most anterior point of the mandible
Gnathion (Gn)	Medial and midpoint between Pog and Me
Menton (Me)	Medial and lowest point of the mandible
Gonion L/R (Go-L / Go-R)	Midpoint of the gonial angle L/R
Infraorbital Foramen L/R (IF-L / IF-R)	External & most distal point of the infraorbital foramen L/R
Internal Acoustic Foramen L/R (IAF-L / IAF-R)	External, most mesial and posterior point of the internal acoustic foramen L/R
Mental Foramen L/R (MF-L / MF-R)	External & most mesial point of the mental foramen L/R
Porion L/R (Po-L / Po-R)	External & uppermost point of the auditory canal L/R
Posterior Nasal Spine (PNS)	Medial & most distal point of the osseous palate
Sella (S)	Central point of the sella

Supplementary Table D3. Definition of the "dental" landmarks localized in our study (FDI World Dental Federation notation for teeth numbering)

Landmark name	Description
11, 21, 31, 41 edges (11E, 21E, 31E, 41E)	Midpoint of 11/21/31/41 incisal edges
11, 21, 31, 41 apexes (11A, 21A, 31A, 41A)	Root apex of 11/21/31/41
16, 26 occlusal (16O, 26O)	Summit of the mesiopalatal cusp of 16/26
36, 46 occlusal (36O, 46O)	Central fossa of 36/46

All CT scans were manually annotated by the same operator (operator #1, a trained orthodontist with at least 5 years of clinical experience). 20 CT scans from the test set were manually annotated a second time by operator #1 and two times by operator #2 (a trained orthodontist with at least 5 years of clinical experience) and operator #3 (a final year postgraduate maxillofacial surgeon) following the procedure described in a previously published repeatability and reproducibility (R&R) study (Dot et al. 2021). The operators only had access to the CT data and segmentations, without any additional clinical information or pre-annotation, and had access neither to each other's results, nor to the index test results, nor to their first session results when performing the second session. Results were exported as an .xml file containing the x-, y-, z- coordinates of each landmark. The ground truth data used to train and test our deep learning model were the annotations of operator #1 for the CT scans landmarked once (*n* = 178), and the means of the 6 annotations by operators #1, #2, #3 for the CT scans landmarked several times (*n* = 20).

In the test set, some CT scans showed missing dental landmarks: 16O (*n* = 1), 26O (*n* = 1), 31A (*n* = 2), 31E (*n* = 2), 36O (*n* = 1), 46O (*n* = 2).

Deep Learning-Based Landmarking (Index Test)

All experiments were performed using the publicly-available SCN framework[1] running in Tensorflow v1.15.0 on our laboratory workstation (CPU AMD Ryzen 9 3900X 12-Core; 128 Gb RAM; GPU Nvidia Titan RTX 24Gb). All our trainings were performed on random image patches of size 128*128*192 voxels, mini-batch size of 1, learning rate 3.10^{-9} and momentum 0.99 (Nesterov's Accelerated Gradient) for 150,000 epochs. Data augmentation was performed on the fly using built-in methods detailed in (Payer et al. 2019). In order to predict each landmark coordinate, we detected the local heatmap maxima for each predicted volume heatmap. Accuracy metrics (distance between reference and predicted landmarks) on the validation set were used to select the final model. Based on the value of the local heatmap maxima, the confidence in a network prediction was considered "very low" when the value was below a threshold (0.4) established from the validation results.

Six networks were trained on our training set (*n* = 128) and evaluated on our validation set (*n* = 32):

- SCN#1 was trained on full scans with a "coarse" resolution of 0.90*0.90*0.62mm^3;
- SCN#2 to SCN#6 were trained on selected regions of interest (ROI) with a "fine" resolution of 0.45*0.45*0.31mm^3 (for orbitale, upper skull base, teeth & anterior maxilla and mandible regions)

[1] https://github.com/christianpayer/MedicalDataAugmentationTool-HeatmapRegression/tree/master/spine (accessed 2022 Jan 07)

or 0.68*0.68*0.47mm^3 (for gonion region). These ROIs were created using the localization of the ground-truth landmarks and margins of 10mm (SCN#2 to SCN#5) or 30mm (SCN#6) in the -x, -y and -z directions.

Inference was performed on our test set (*n* = 38) with a 2-stage method (Fig. 1B), following a sliding-window approach on image patches of the same size and resolution as the trained network. At stage 1, SCN#1 was used to predict the "coarse" localization of the landmarks. These results allowed us to extract the 5 ROIs using the localization of the predicted landmarks and margins of 10mm (SCN#2 to SCN#5) or 30mm (SCN#6) in the -x, -y and -z directions. At stage 2, SCN#2 to SCN#6 were used to predict the "fine" localization of the landmarks. Afterwards, the locally-predicted landmarks coordinates were transferred back into the general coordinate system for result evaluation. This method systematically localized 33 landmarks for each CT scan. In CT scans with missing landmarks (*i.e.* missing teeth), the corresponding predictions were considered as missing values and deleted by the operator.

Data preprocessing steps (e.g., normalization, rescaling, cropping of the image patches) were done automatically by the SCN. Other data processing steps (creation of ROIs, coordinate system transformations) were performed automatically using custom-made Python (http://www.python.org, version 3.7) scripts.

Evaluation

Localization performance

Each landmark was subjected to statistical analysis to compare the errors obtained from scans with reference constructed from 1 annotation with the errors obtained from scans with reference constructed from means of 6 annotations.

Cephalometric measurements

A conventional cephalometric analysis (Supplementary Table D4) was conducted; nine 2D angles (degrees) and six 2D distances (mm) were calculated using orthogonal projections of the 3D landmarks on an automatically constructed midsagittal plane (MSP). MSP construction followed two steps: 1) the CT scans were segmented using a previously published DL-based automated method (Dot et al. 2022); 2) the MSP was computed thanks to the upper skull segmentation using a previously published automated method (Pinheiro et al. 2019). For each variable and each CT scan, the difference between the measurements obtained from reference landmarks and from predicted landmarks was computed and the proportion of measurements with differences under 2mm or 2° was calculated. Additionally,

the accuracy of Frankfort horizontal (FH) plane construction (porion right/left and orbitale left) was evaluated by computing the absolute angular distances between reference and predicted FH planes.

Supplementary Table D4. Skeletal and dentoalveolar cephalometric variables used in this study.

Variable	Description
Skeletal	
SNA (°)	Angle between projected line S-Na and projected line Na-A (lateral view)
SNB (°)	Angle between projected line S-Na and projected line Na-B (lateral view)
ANB (°)	Angle between projected line Na-A and projected line Na-B (lateral view)
ANS-PNS / Go-Gn (°)	Angle between projected line ANS-PNS and projected line Mean_Go-Gn (lateral view)
S-Na / Go-Gn (°)	Angle between projected line S-Na and projected line Mean_Go-Gn (lateral view)
Pog to Na-B (mm)	Orthogonal distance between Pog and projected line Na-B (lateral view)
A to MSP (mm)	Orthogonal distance between A and projected MSP (frontal view)
B to MSP (mm)	Orthogonal distance between B and projected MSP (frontal view)
Pog to MSP (mm)	Orthogonal distance between Pog and projected MSP (frontal view)
Dentoalveolar	
S-Na / Occlusal plane (°)	Angle between projected line S-Na and projected occlusal plane (mean_160_26O-mean_11E_21E) (lateral view)
U1 / ANS-PNS (°)	Angle between projected line mean_11E_21E-mean_11A_21A and projected line ANS-PNS (lateral view)
U1 to Na-A (mm)	Orthogonal distance between mean_11E_21E and projected line Na-A (lateral view)
Interincisal angle (°)	Angle between projected line mean_11E_21E-mean_11A_21A and projected line mean_31E_41E-mean_31A_41A (lateral view)
L1 / Go-Gn (°)	Angle between projected line mean_31E_41E-mean_31A_41A and projected line Go-Gn (lateral view)
L1 to Na-B (mm)	Orthogonal distance between mean_11E_21E and projected line Na-B (lateral view)

Comparison with manual landmarking reproducibility

The results from a previous R&R study were used to assess the Bland-Altman 95% limits of agreement (LoA) of manual landmarking reproducibility (Dot et al. 2021). The proportion of predicted landmarks within these limits was computed for each *-x -y -z* axis. In addition, we applied ISO norm 5725 on the cephalometric variables to calculate the 95% LoA of manual measurement reproducibility (Supplementary Table D5) (ISO 5725-2:2019). The proportion of predicted cephalometric variables within these limits was computed. For the CT scans included in the R&R study, statistical tests were used to compare automatic and manual results, and boxplots of the localization and measurement errors were computed.

Supplementary Table D5. Bland-Altman 95% limits of agreement (LoA) of manual repeatability (2 repetitions) and reproducibility (3 operators) for the cephalometric variables (°/mm), calculated on 20 CT scans following the ISO 5725 standard.

	Repeatability 95% LoA	**Reproducibility 95% LoA**
Skeletal		
SNA (°)	1.029	1.319
SNB (°)	0.975	1.318
ANB (°)	0.414	0.456
ANS-PNS / Go-Gn (°)	1.601	2.252
S-Na / Go-Gn (°)	1.621	1.979
Pog to Na-B (mm)	0.645	0.791
A to MSP (mm)	0.802	0.865
B to MSP (mm)	0.648	1.123
Pog to MSP (mm)	0.712	1.142
Dentoalveolar		
S-Na / Occlusal plane (°)	0.826	1.107
U1 / ANS-PNS (°)	1.53	2.148
U1 to Na-A (mm)	0.405	0.464
Interincisal angle (°)	1.474	2.127
L1 / Go-Gn (°)	1.306	1.994
L1 to Na-B (mm)	0.38	0.504

Results

Localization performance

Supplementary Table D6. Mean radial errors (mm), success detection rates (% (*n*)) and minimum/maximum radial error (mm) for each landmark on the test set with the outlier case included (*n* = 38). MRE, mean radial error; SD, standard deviation; Min., minimum radial error; Max., maximum radial error; L, left; R, right.

	MRE ± SD	<2mm	<2.5mm	<3mm	Min.	Max.
11 Apex	1.7 ± 6.2	97.4 (37)	97.4 (37)	97.4 (37)	0.2	39.0
11 Edge	0.6 ± 1.0	97.4 (37)	97.4 (37)	97.4 (37)	0.1	6.2
16 Occlusal	1.3 ± 2.4	94.6 (35)	94.6 (35)	94.6 (35)	0.1	11.2
21 Apex	0.7 ± 0.4	100 (38)	100 (38)	100 (38)	0.2	1.9
21 Edge	0.7 ± 1.2	97.4 (37)	97.4 (37)	97.4 (37)	0.1	7.8
26 Occlusal	1.6 ± 3.5	91.9 (34)	91.9 (34)	91.9 (34)	0.1	16.6
31 Apex	1.5 ± 3.8	94.4 (34)	94.4 (34)	94.4 (34)	0.2	22.2
31 Edge	1.1 ± 3.3	91.7 (33)	94.4 (34)	94.4 (34)	0.1	18.9
36 Occlusal	1.8 ± 3.4	89.2 (33)	89.2 (33)	89.2 (33)	0.2	12.7
41 Apex	1.1 ± 2.9	97.4 (37)	97.4 (37)	97.4 (37)	0.2	18.3
41 Edge	0.6 ± 1.1	97.4 (37)	97.4 (37)	97.4 (37)	0.1	7.1
46 Occlusal	0.9 ± 1.8	97.2 (35)	97.2 (35)	97.2 (35)	0.1	11.0
A Point	1.6 ± 2.9	86.9 (33)	89.5 (34)	89.5 (34)	0.2	18.1
Anterior Nasal Spine	0.7 ± 0.7	94.7 (36)	94.8 (36)	97.4 (37)	0.1	3.2
B Point	1.7 ± 1.5	65.8 (25)	81.6 (31)	92.1 (35)	0.3	8.5
Gnathion	1.0 ± 0.6	92.1 (35)	97.4 (37)	100 (38)	0.3	2.5
Gonion L	1.9 ± 1.7	68.4 (26)	76.3 (29)	86.8 (33)	0.3	7.3
Gonion R	2.1 ± 1.4	50 (19)	71.1 (27)	73.7 (28)	0.3	6.8
Infraorbital Foramen L	0.6 ± 0.3	100 (38)	100 (38)	100 (38)	0.2	2.0
Infraorbital Foramen R	0.6 ± 0.5	97.4 (37)	100 (38)	100 (38)	0.1	2.4
Internal Acoustic Foramen L	0.6 ± 0.4	100 (38)	100 (38)	100 (38)	0.2	1.9
Internal Acoustic Foramen R	0.6 ± 0.6	97.4 (37)	97.4 (37)	97.4 (37)	0.1	3.9
Mental Foramen L	0.4 ± 0.2	100 (38)	100 (38)	100 (38)	0.1	0.8
Mental Foramen R	0.4 ± 0.3	100 (38)	100 (38)	100 (38)	0.1	1.3
Menton	1.0 ± 0.6	94.7 (36)	97.4 (37)	100 (38)	0.4	2.6
Nasion	0.7 ± 0.4	100 (38)	100 (38)	100 (38)	0.1	1.9
Orbitale L	2.6 ± 2.0	44.7 (17)	57.9 (22)	68.4 (26)	0.1	8.8
Orbitale R	2.6 ± 2.3	55.3 (21)	65.8 (25)	68.4 (26)	0.3	9.7
Pogonion	1.1 ± 0.6	89.5 (34)	97.4 (37)	100 (38)	0.2	3.0
Porion L	1.1 ± 0.5	89.5 (34)	100 (38)	100 (38)	0.2	2.3
Porion R	1.3 ± 0.7	86.8 (33)	89.5 (34)	100 (38)	0.3	2.8
Posterior Nasal Spine	0.5 ± 0.4	100 (38)	100 (38)	100 (38)	0.1	1.5
Sella	0.8 ± 0.4	100 (38)	100 (38)	100 (38)	0.2	2.0

Supplementary Table D7. Radial errors (mm) for each landmark of the outlier case. L, left; R, right.

	Euclidian Error (mm)
11 Apex	39.0
11 Edge	6.2
16 Occlusal	1.4
21 Apex	1.3
21 Edge	7.8
26 Occlusal	16.6
31 Apex	22.2
31 Edge	18.9
36 Occlusal	12.7
41 Apex	18.3
41 Edge	7.1
46 Occlusal	
A Point	18.1
Anterior Nasal Spine	0.3
B Point	2.2
Gnathion	0.4
Gonion L	2.1
Gonion R	0.8
Infraorbital Foramen L	0.5
Infraorbital Foramen R	0.5
Internal Acoustic Foramen L	0.2
Internal Acoustic Foramen R	0.5
Mental Foramen L	0.5
Mental Foramen R	0.7
Menton	0.9
Nasion	1.5
Orbitale L	0.2
Orbitale R	3.6
Pogonion	1.0
Porion L	1.1
Porion R	0.8
Posterior Nasal Spine	0.5
Sella	0.9

Supplementary Table D8. Mean radial errors (mm), success detection rates (% (*n*)) and minimum/maximum radial error (mm) for each landmark on the validation set (*n* = 32). MRE, mean radial error; SD, standard deviation; Min., minimum radial error; Max., maximum radial error; L, left; R, right.

	MRE ± SD	<2mm	<2.5mm	<3mm	Min.	Max.
11 Apex	0.7 ± 0.3	100 (31)	100 (31)	100 (31)	0.3	1.6
11 Edge	0.5 ± 0.3	100 (31)	100 (31)	100 (31)	0.1	1.2
16 Occlusal	1.4 ± 2.3	83.9 (26)	90.3 (28)	90.3 (28)	0.2	12.8
21 Apex	0.7 ± 0.3	100 (30)	100 (30)	100 (30)	0.2	1.3
21 Edge	0.4 ± 0.2	100 (30)	100 (30)	100 (30)	0.1	1.2
26 Occlusal	1.3 ± 2.4	90 (27)	93.3 (28)	93.3 (28)	0.3	13.4
31 Apex	0.7 ± 0.4	100 (32)	100 (32)	100 (32)	0.1	2.0
31 Edge	0.5 ± 0.2	100 (32)	100 (32)	100 (32)	0.2	1.0
36 Occlusal	1.0 ± 1.9	96.4 (27)	96.4 (27)	96.4 (27)	0.2	10.4
41 Apex	0.7 ± 0.4	96.9 (31)	96.9 (31)	100 (32)	0.2	2.6
41 Edge	0.4 ± 0.2	100 (32)	100 (32)	100 (32)	0.1	0.9
46 Occlusal	1.2 ± 2.3	93.3 (28)	96.7 (29)	96.7 (29)	0.2	13.1
A Point	1.1 ± 0.8	90.6 (29)	93.8 (30)	96.9 (31)	0.3	4.0
Anterior Nasal Spine	0.9 ± 0.7	93.8 (30)	93.8 (30)	96.9 (31)	0.1	3.3
B Point	1.9 ± 1.3	59.4 (19)	68.8 (22)	78.1 (25)	0.3	5.0
Gnathion	1.0 ± 0.5	96.9 (31)	100 (32)	100 (32)	0.2	2.3
Gonion L	1.4 ± 1.0	84.4 (27)	90.6 (29)	90.6 (29)	0.2	4.8
Gonion R	1.5 ± 1.0	71.9 (23)	78.1 (25)	90.6 (29)	0.2	3.8
Infraorbital Foramen L	0.8 ± 0.5	96.9 (31)	100 (32)	100 (32)	0.1	2.3
Infraorbital Foramen R	0.9 ± 0.6	93.8 (30)	96.9 (31)	100 (32)	0.2	2.8
Internal Acoustic Foramen L	0.5 ± 0.4	100 (32)	100 (32)	100 (32)	0.2	1.6
Internal Acoustic Foramen R	0.7 ± 0.4	100 (32)	100 (32)	100 (32)	0.1	1.9
Mental Foramen L	0.5 ± 0.5	96.9 (31)	96.9 (31)	100 (32)	0.1	2.8
Mental Foramen R	0.4 ± 0.2	100 (32)	100 (32)	100 (32)	0.1	1.0
Menton	1.1 ± 0.6	96.9 (31)	96.9 (31)	100 (32)	0.4	2.9
Nasion	0.7 ± 0.4	100 (32)	100 (32)	100 (32)	0.1	1.6
Orbitale L	1.8 ± 1.4	75 (24)	78.1 (25)	78.1 (25)	0.2	5.1
Orbitale R	1.6 ± 1.1	75 (24)	75 (24)	90.6 (29)	0.2	4.6
Pogonion	1.1 ± 0.6	93.8 (30)	96.9 (31)	100 (32)	0.3	2.6
Porion L	1.1 ± 0.8	84.4 (27)	90.6 (29)	96.9 (31)	0.3	3.1
Porion R	1.1 ± 0.6	90.6 (29)	100 (32)	100 (32)	0.2	2.3
Posterior Nasal Spine	0.8 ± 1.2	93.8 (30)	96.9 (31)	96.9 (31)	0.1	6.8
Sella	0.8 ± 0.4	100 (32)	100 (32)	100 (32)	0.3	1.6

Comparison with manual landmarking and measurement reproducibility

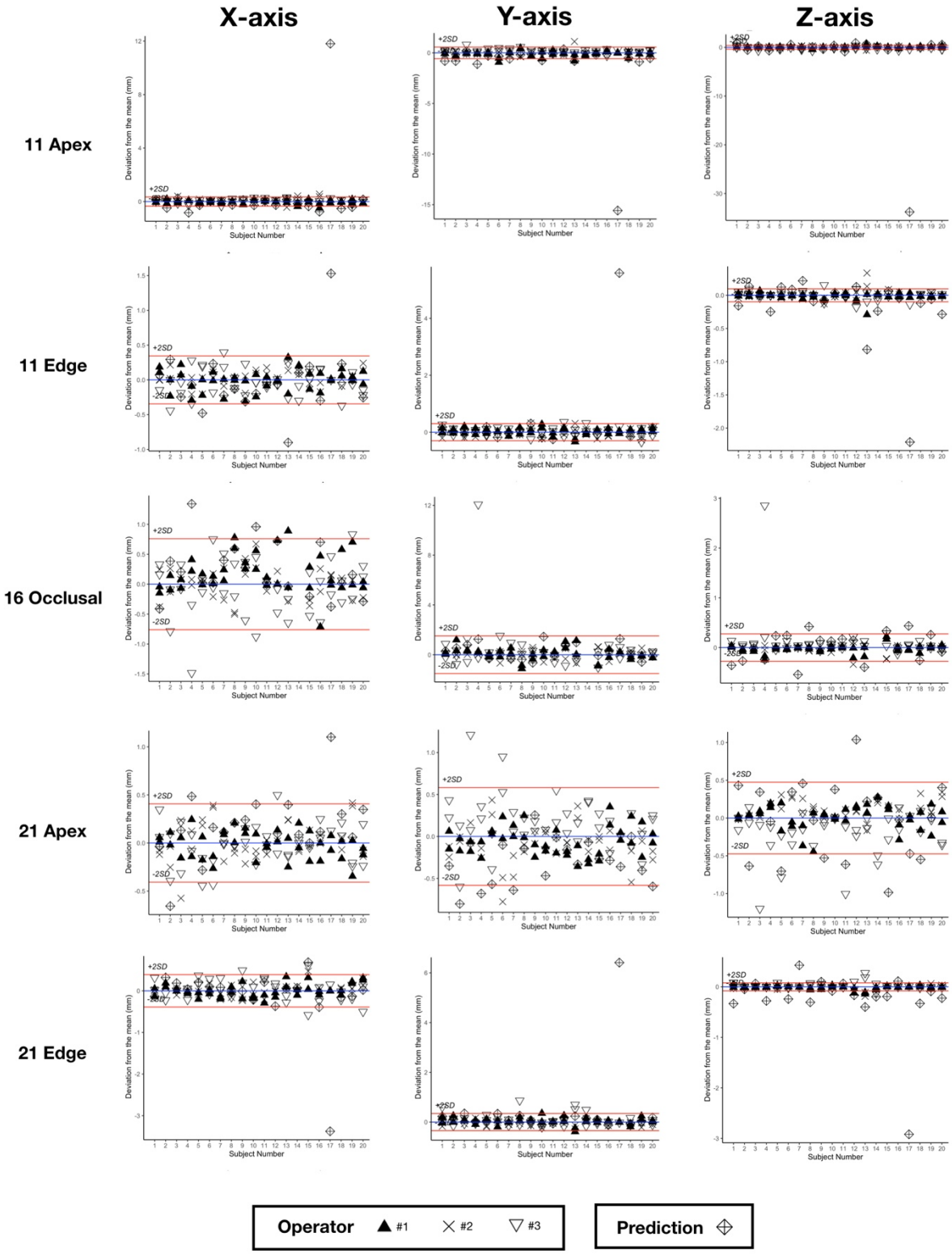
Supplementary Table D9. Proportion (% (*n*)) of predicted landmarks coordinates in the -x, -y and -z directions within Bland-Altman 95% limits of agreement of manual reproducibility (95% LoA) on the hold-out test set without the outlier case (*n* = 37). L, left; R, right.

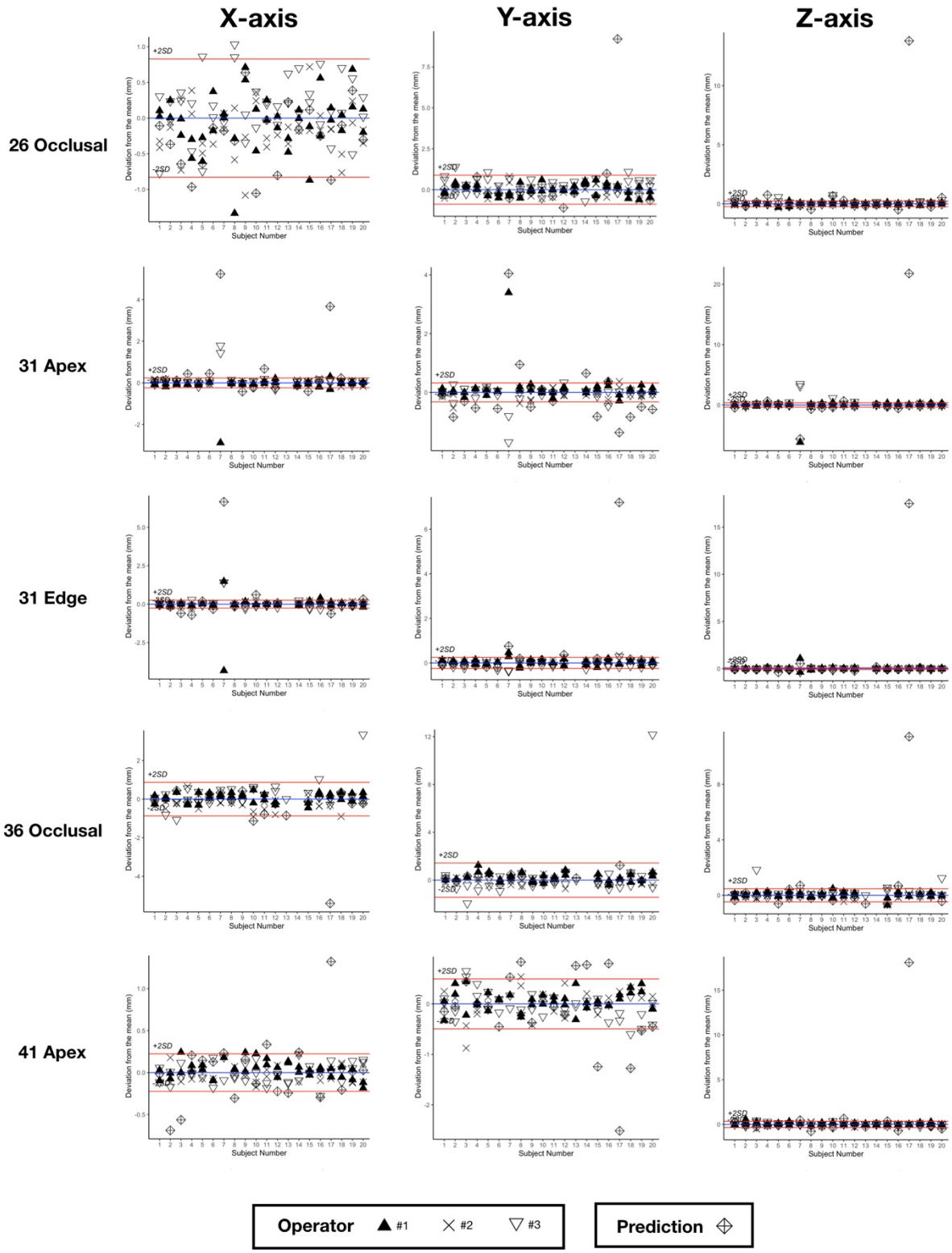
	Within 95% LoA		
	X Axis	Y Axis	Z Axis
11 Apex	67.6 (25)	67.6 (25)	62.2 (23)
11 Edge	73 (27)	81.1 (30)	40.5 (15)
16 Occlusal	81.1 (30)	91.9 (34)	56.8 (21)
21 Apex	86.5 (32)	75.7 (28)	64.9 (24)
21 Edge	73 (27)	83.8 (31)	16.2 (6)
26 Occlusal	89.2 (33)	86.5 (32)	56.8 (21)
31 Apex	59.5 (22)	27 (10)	54.1 (20)
31 Edge	59.5 (22)	64.9 (24)	35.1 (13)
36 Occlusal	86.5 (32)	89.2 (33)	64.9 (24)
41 Apex	51.4 (19)	64.9 (24)	70.3 (26)
41 Edge	56.8 (21)	56.8 (21)	35.1 (13)
46 Occlusal	78.4 (29)	81.1 (30)	64.9 (24)
A Point	100 (37)	78.4 (29)	83.8 (31)
Anterior Nasal Spine	91.9 (34)	91.9 (34)	97.3 (36)
B Point	91.9 (34)	94.6 (35)	91.9 (34)
Gnathion	86.5 (32)	89.2 (33)	91.9 (34)
Gonion L	89.2 (33)	86.5 (32)	83.8 (31)
Gonion R	97.3 (36)	70.3 (26)	78.4 (29)
Infraorbital Foramen L	97.3 (36)	100 (37)	100 (37)
Infraorbital Foramen R	94.6 (35)	94.6 (35)	91.9 (34)
Internal Acoustic Foramen L	97.3 (36)	94.6 (35)	100 (37)
Internal Acoustic Foramen R	94.6 (35)	97.3 (36)	100 (37)
Mental Foramen L	89.2 (33)	89.2 (33)	94.6 (35)
Mental Foramen R	94.6 (35)	86.5 (32)	75.7 (28)
Menton	89.2 (33)	100 (37)	91.9 (34)
Nasion	83.8 (31)	67.6 (25)	94.6 (35)
Orbitale L	75.7 (28)	86.5 (32)	86.5 (32)
Orbitale R	73 (27)	89.2 (33)	86.5 (32)
Pogonion	94.6 (35)	89.2 (33)	94.6 (35)
Porion L	100 (37)	83.8 (31)	94.6 (35)
Porion R	100 (37)	94.6 (35)	89.2 (33)
Posterior Nasal Spine	97.3 (36)	91.9 (34)	91.9 (34)
Sella	91.9 (34)	91.9 (34)	97.3 (36)

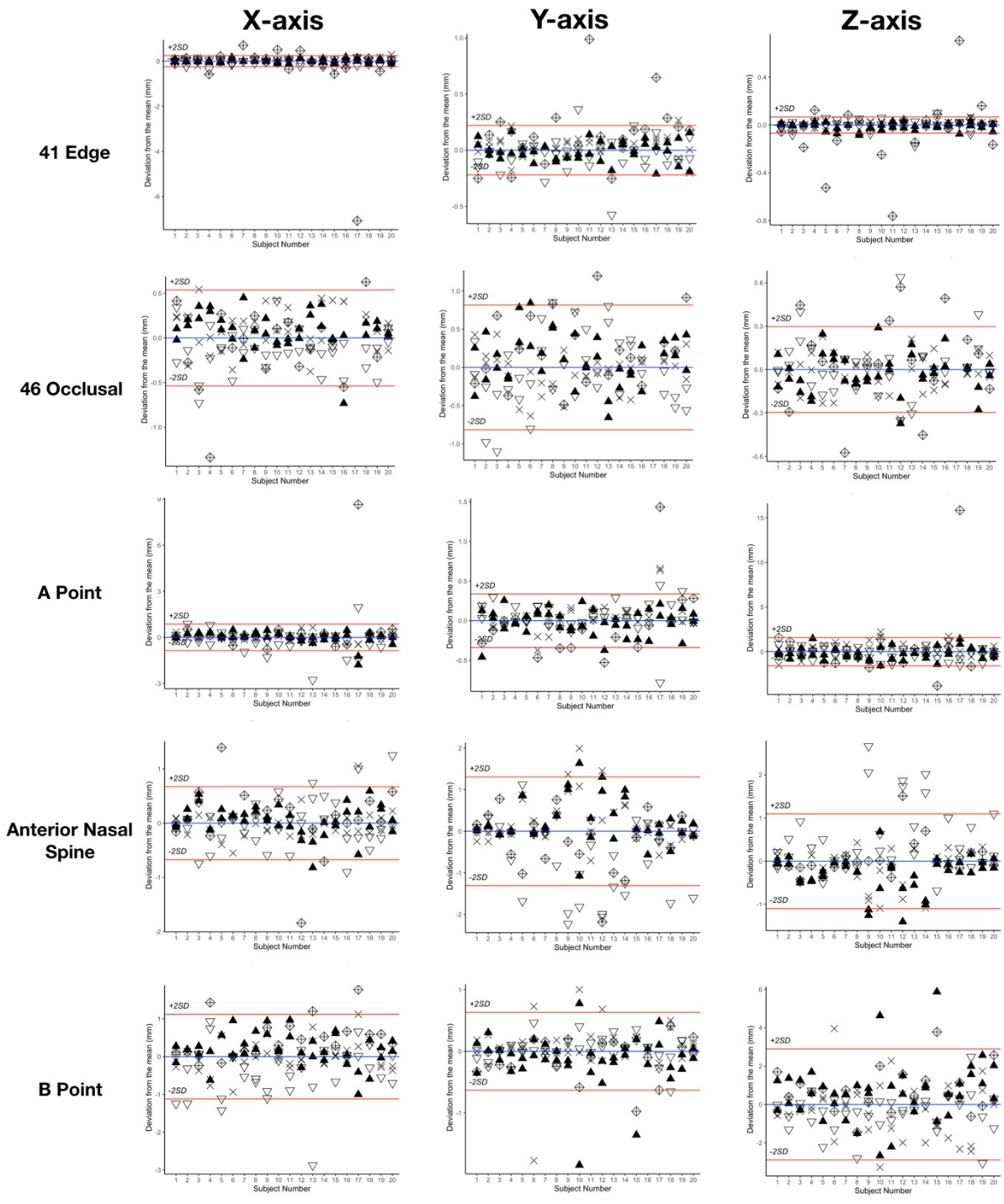
Supplementary Table D10. Proportion (% (*n*)) of predicted cephalometric measurements within Bland-Altman 95% limits of agreement of manual reproducibility (95% LoA) on the hold-out test set without the outlier case (*n* = 37).

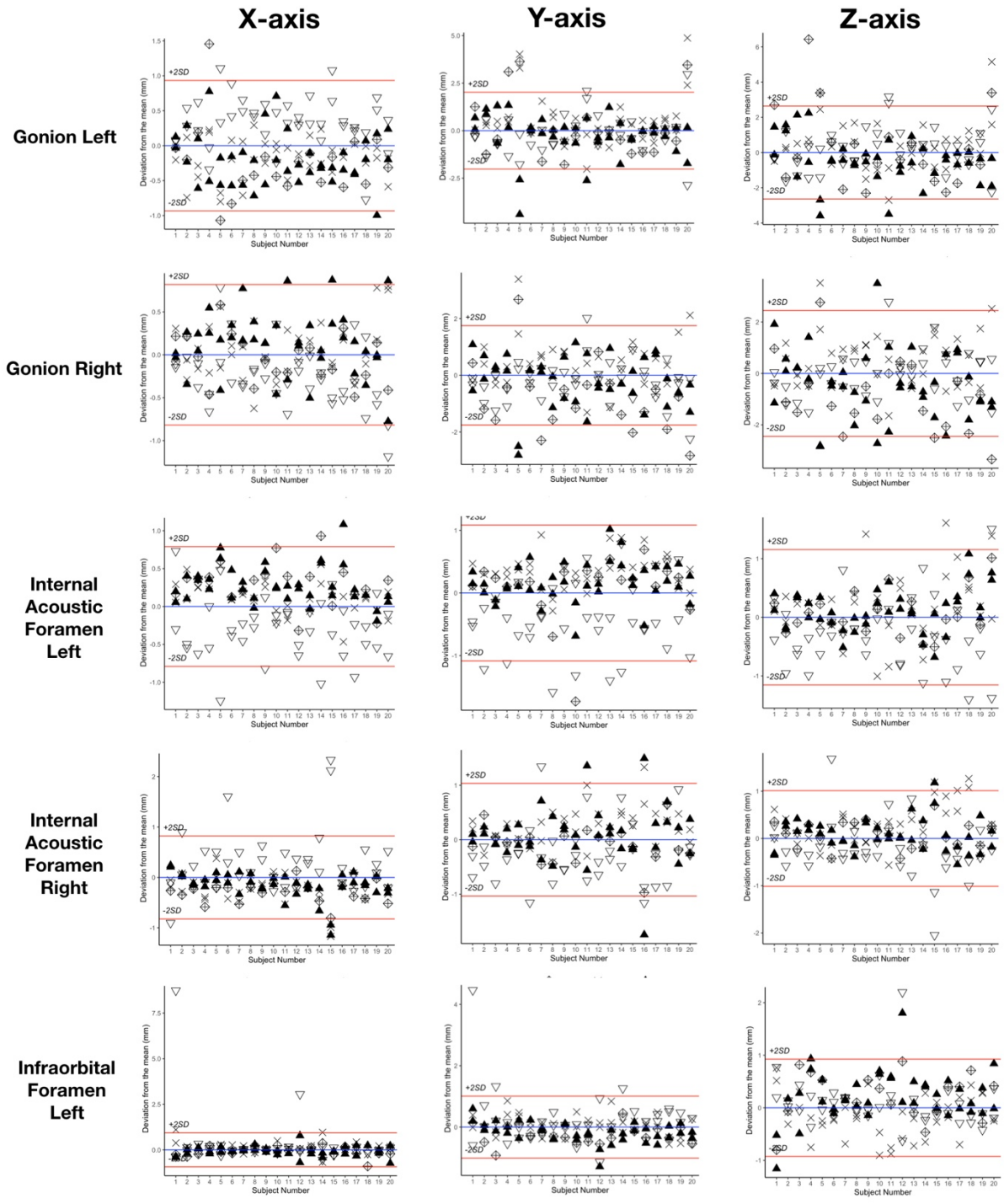
	Within 95% LoA
Skeletal	
SNA (°)	86.5 (32)
SNB (°)	91.9 (34)
ANB (°)	91.9 (34)
ANS-PNS / Go-Gn (°)	94.6 (35)
S-Na / Go-Gn (°)	81.1 (30)
Pog to Na-B (mm)	94.6 (35)
A to MSP (mm)	100 (37)
B to MSP (mm)	91.9 (34)
Pog to MSP (mm)	94.6 (35)
Dentoalveolar	
S-Na / Occlusal plane (°)	62.9 (22)
U1 / ANS-PNS (°)	78.4 (29)
U1 to Na-A (mm)	75.7 (28)
Interincisal angle (°)	57.1 (20)
L1 / Go-Gn (°)	77.1 (27)
L1 to Na-B (mm)	78.4 (29)

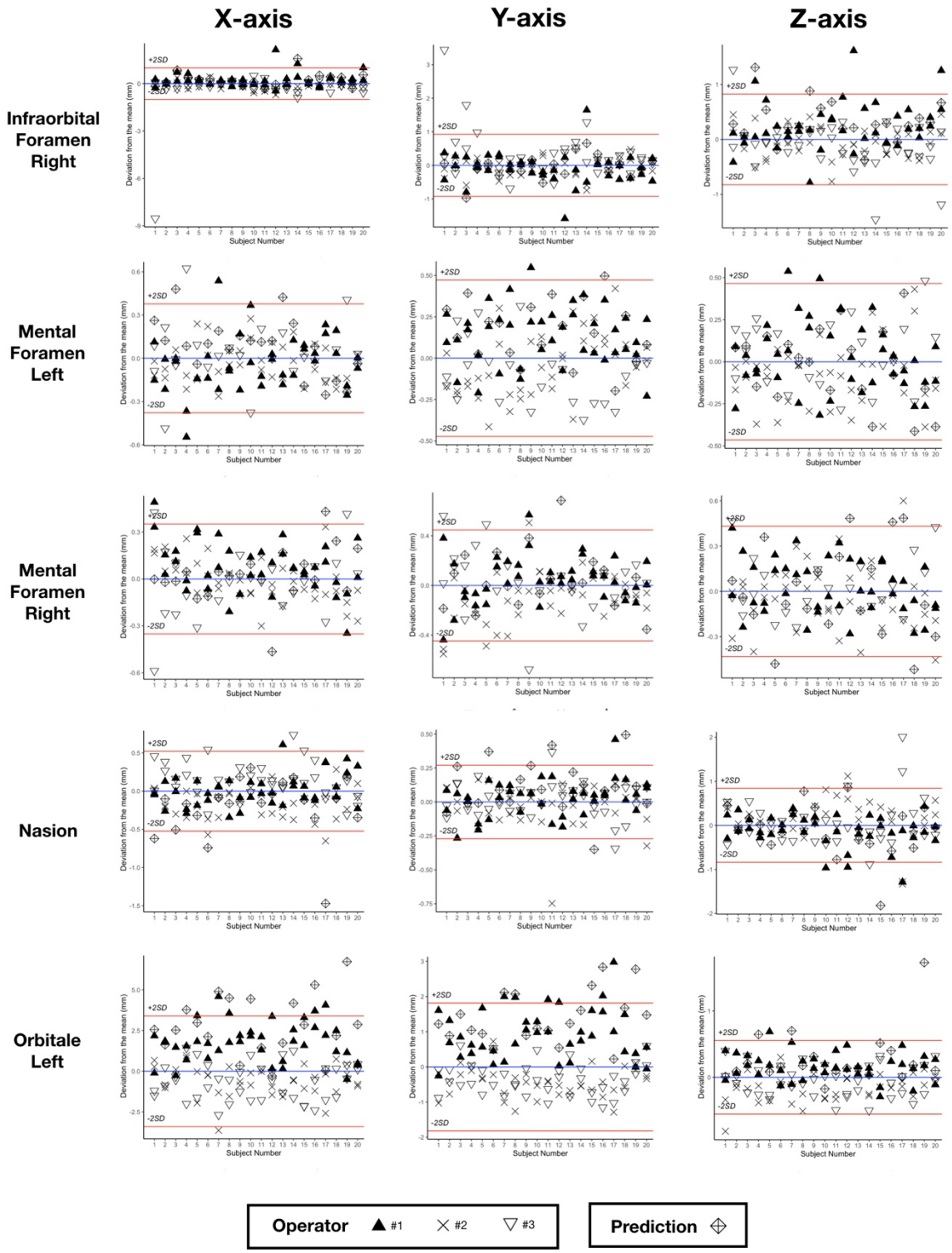
Bland-Altman plots of landmarking localization. For the 33 landmarks, the following plots show the deviations from the mean (blue line) of the 6 manual repetitions and the predictions for the 20 subjects included in the R&R study. Please note that the scales differ. Subject number 17 is the outlier case. Red lines show the ± 2×SD of reproducibility. SD, standard deviation.

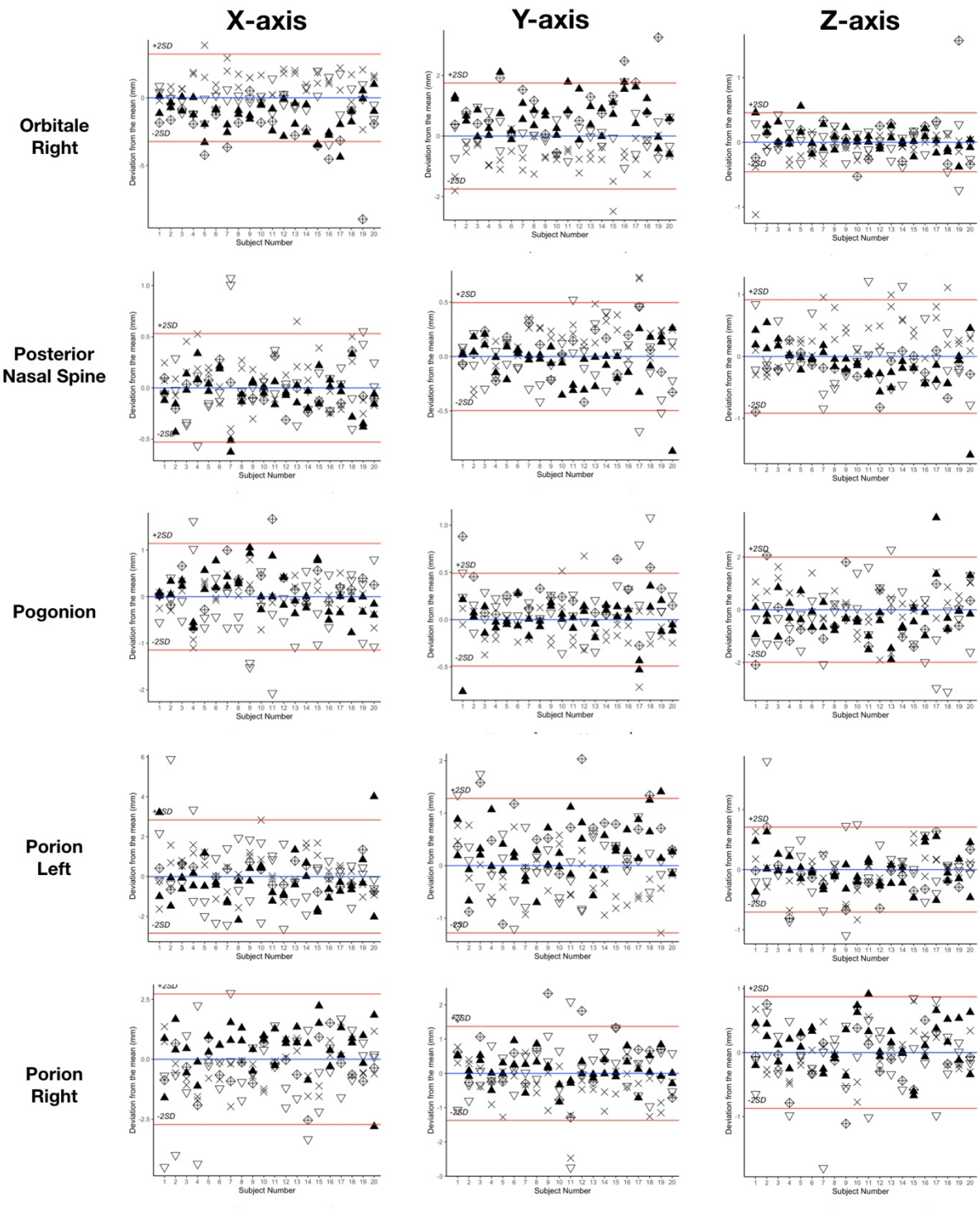


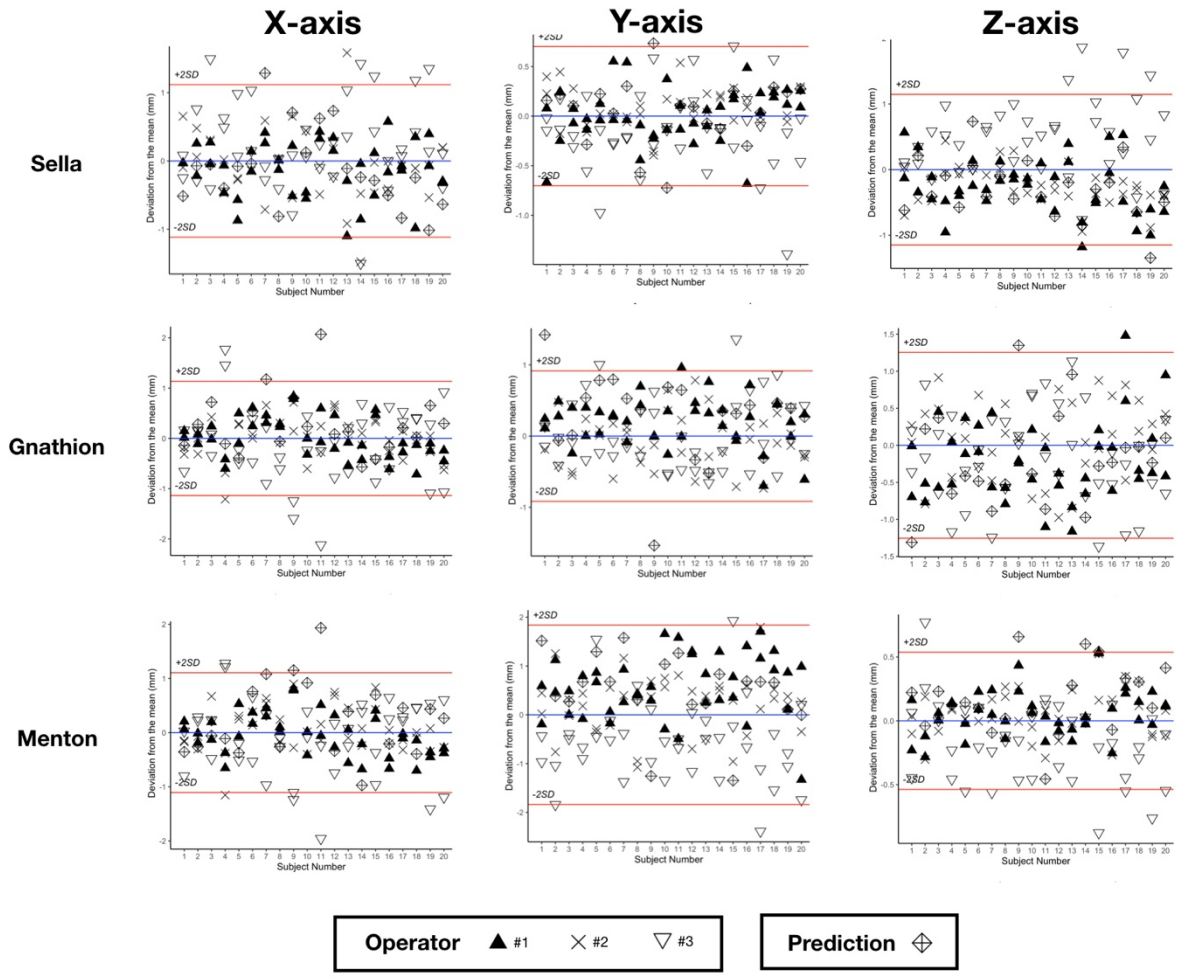

X-axis
Y-axis
Z-axis
26 Occlusal
31 Apex
31 Edge
36 Occlusal
41 Apex
Deviation from the mean (mm)
Subject Number
+2SD
-2SD
Operator
#1
#2
#3
Prediction


X-axis
Y-axis
Z-axis
41 Edge
46 Occlusal
A Point
Anterior Nasal Spine
B Point
Deviation from the mean (mm)
Subject Number
+2SD
-2SD
Operator
#1
#2
#3
Prediction

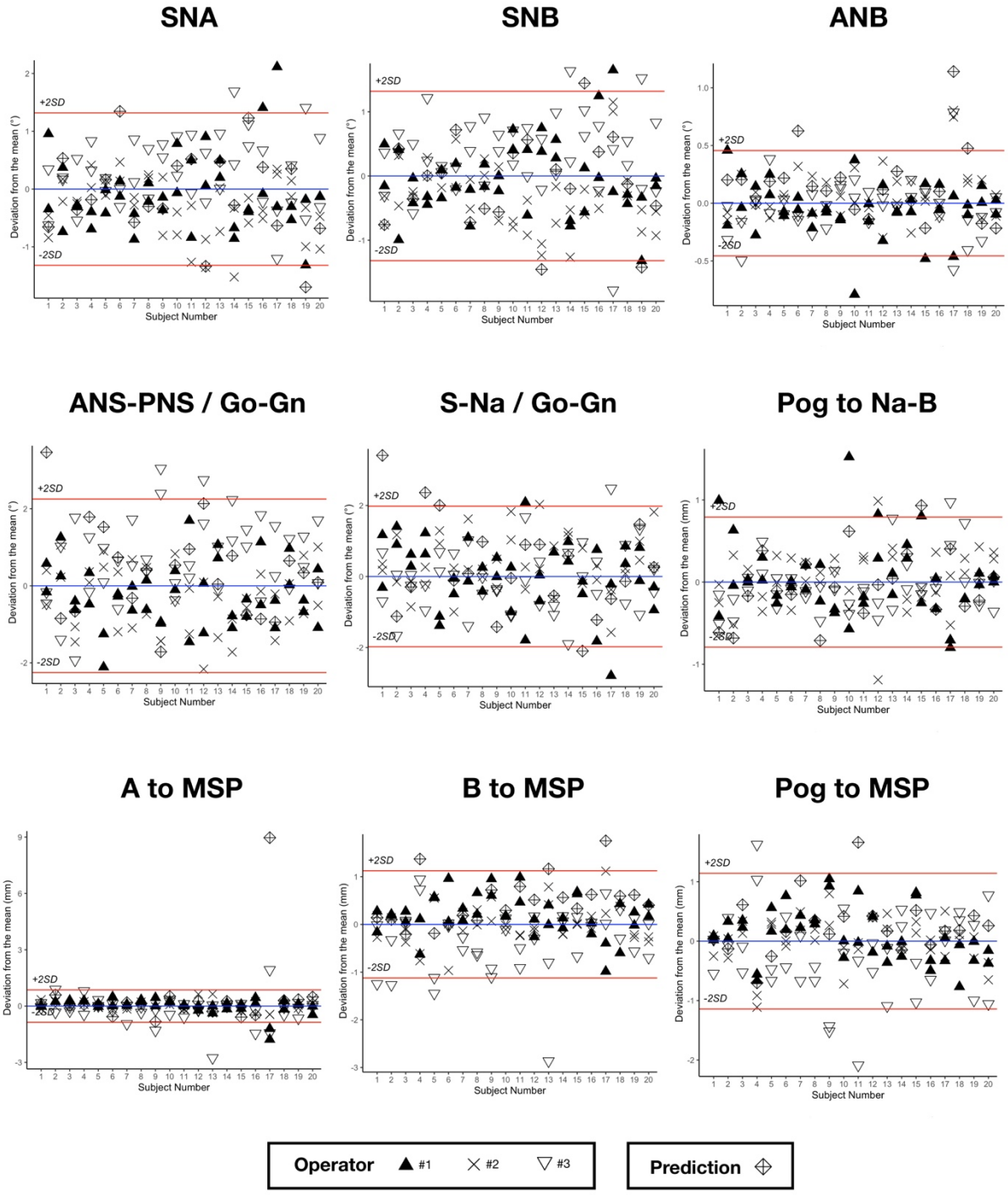

X-axis
Y-axis
Z-axis
Gonion Left
Gonion Right
Internal Acoustic Foramen Left
Internal Acoustic Foramen Right
Infraorbital Foramen Left
Deviation from the mean (mm)
Subject Number
+2SD
-2SD
Operator
#1
#2
#3
Prediction

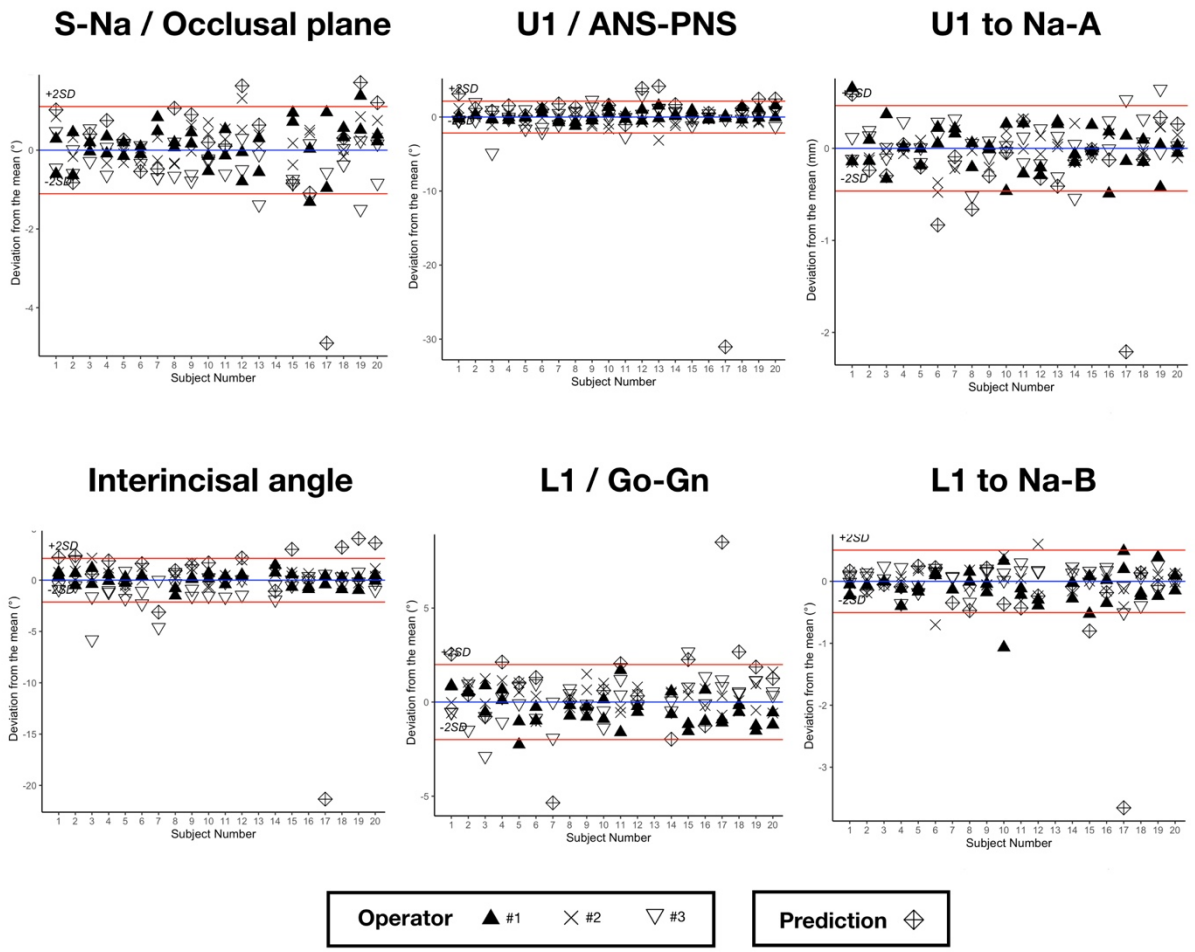

X-axis
Y-axis
Z-axis
Infraorbital Foramen Right
Mental Foramen Left
Mental Foramen Right
Nasion
Orbitale Left
+2SD
-2SD
Deviation from the mean (mm)
Subject Number
Operator
#1
#2
#3
Prediction


X-axis
Y-axis
Z-axis
Orbitale Right
Posterior Nasal Spine
Pogonion
Porion Left
Porion Right
+2SD
-2SD
Deviation from the mean (mm)
Subject Number
Operator
#1
#2
#3
Prediction


X-axis
Y-axis
Z-axis
Sella
Gnathion
Menton
+2SD
-2SD
Deviation from the mean (mm)
Subject Number
Operator
#1
#2
#3
Prediction

Bland-Altman plots of cephalometric measurements. For the 15 measurements, the following plots show the deviations from the mean (blue line) of the 6 manual repetitions and the predictions for the 20 subjects included in the R&R study. Please note that the scales differ. Subject number 17 is the outlier case. Red lines show the ± 2×SD of reproducibility. SD, standard deviation.




S-Na / Occlusal plane
U1 / ANS-PNS
U1 to Na-A
Interincisal angle
L1 / Go-Gn
L1 to Na-B
+2SD
-2SD
Deviation from the mean (°)
Deviation from the mean (mm)
Subject Number
Operator
#1
#2
#3
Prediction

HESAM UNIVERSITÉ

Gauthier DOT

Arts et Métiers Sciences et Technologies

Automatisation du traitement des imageries tridimensionnelles dento-maxillo-faciales par apprentissage profond : application à la segmentation et à la céphalométrie

Résumé : L'utilisation clinique d'imageries tridimensionnelles (3D) dento-maxillo-faciales s'est fortement développée ces dernières années, permettant d'améliorer le diagnostic et la planification de certains traitements orthodontiques et orthodontico-chirurgicaux. Le traitement de ces imageries 3D est cependant contraignant et repose sur de nombreuses étapes manuelles, nécessitant plusieurs niveaux de validation, du temps et des opérateurs formés. La routine clinique reste largement basée sur l'utilisation de méthodes 2D, peu adaptées pour les patients présentant des dysmorphies faciales complexes comme des asymétries importantes ou des syndromes cranio-faciaux.

L'objectif principal de ce travail a été de mettre en œuvre des modèles d'apprentissage profond afin d'automatiser deux étapes du traitement de ces imageries 3D : (1) la reconstruction des modèles surfaciques 3D, processus appelé « segmentation » et (2) le placement de points d'intérêts anatomiques pour la réalisation d'une analyse céphalométrique 3D. L'évaluation de ces modèles a été effectuée sur une base de données originale de patients présentant des dysmorphies faciales variées et marquées, en comparant la performance de l'algorithme avec celle d'experts sur la base de critères présentant une pertinence clinique.

Sur une base de données de test de 153 scanners, la segmentation automatisée a présenté un coefficient de Dice surfacique à 1 mm de 98.03 ± 2.48 %, 148 scanners présentant un score moyen supérieur au seuil de viabilité clinique de 95 %. Sur une base de données de test de 37 scanners, l'erreur moyenne du placement des points céphalométriques était de 1.0 ± 1.3 mm et 90.4 % des prédictions étaient situées à moins de 2 mm de la référence. Une validation plus large, incluant des données d'autres centres cliniques, devra être effectuée afin d'évaluer le potentiel de généralisation de ces résultats. Trois cas cliniques sont présentés pour illustrer les perspectives d'applications cliniques de ces résultats.

Mots clés : tomodensitométrie ; interprétation d'images radiographiques assistée par ordinateur ; céphalométrie ; repères anatomiques ; apprentissage profond ; intelligence artificielle

Automation of three-dimensional dentomaxillofacial image processing via deep learning: application to segmentation and cephalometry

Abstract : The clinical use of three-dimensional (3D) dentomaxillofacial imaging has developed significantly in recent years, allowing for improved diagnosis and planning of some orthodontic and orthodontic-surgical treatments. However, the processing of these 3D images remains restrictive and relies on many manual steps, requiring several levels of validation, time and trained operators. The clinical routine is still largely based on the use of 2D methods, which are not well adapted for patients with complex facial deformities such as important asymmetries or craniofacial syndromes.

The main objective of this work was to implement deep learning models in order to automate two steps in the processing of these 3D images: (1) the reconstruction of 3D surface models, a process called "segmentation", and (2) the placement of anatomical landmarks for 3D cephalometric analysis. The evaluation of these models was performed on an original database of patients with varied and marked facial deformities, comparing the performance of the algorithm with that of experts on the basis of clinically relevant metrics.

In a test database of 153 CT scans, the automated segmentation had a surface Dice Similarity Coefficient at 1mm of 98.03 ± 2.48%, with 148 scans having a mean score which cleared the 95% limit for clinical significance. In a test database of 37 scans, the mean error of cephalometric landmark localization was 1.0 ± 1.3 mm, and 90.4% of predictions were within 2 mm of the reference. A broader validation, including data from other clinical centers, will need to be performed to assess the generalizability of these results. Three clinical cases illustrate the perspectives of clinical applications of these results.

Keywords: tomography, x-ray computed; radiographic image interpretation, computer-assisted; cephalometry; anatomic landmarks; deep learning; artificial intelligence