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 Paristech, Paris, France ;

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

[5] EPF, Graduate School of Engineering, Cachan, France.

* Corresponding author

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.

Keywords: Orthodontics; Orthognathic Surgery; Surgery, Computer-Assisted; Artificial Intelligence; Tomography, X-ray Computed; Cephalometry; Anatomic Landmarks

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 (American Academy of Oral and Maxillofacial Radiology 2013; Kapila and Nervina 2015). 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 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.

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).

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) (Appendix Figure 1). 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 Appendix Table 1.

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 appendix.

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 appendix 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.

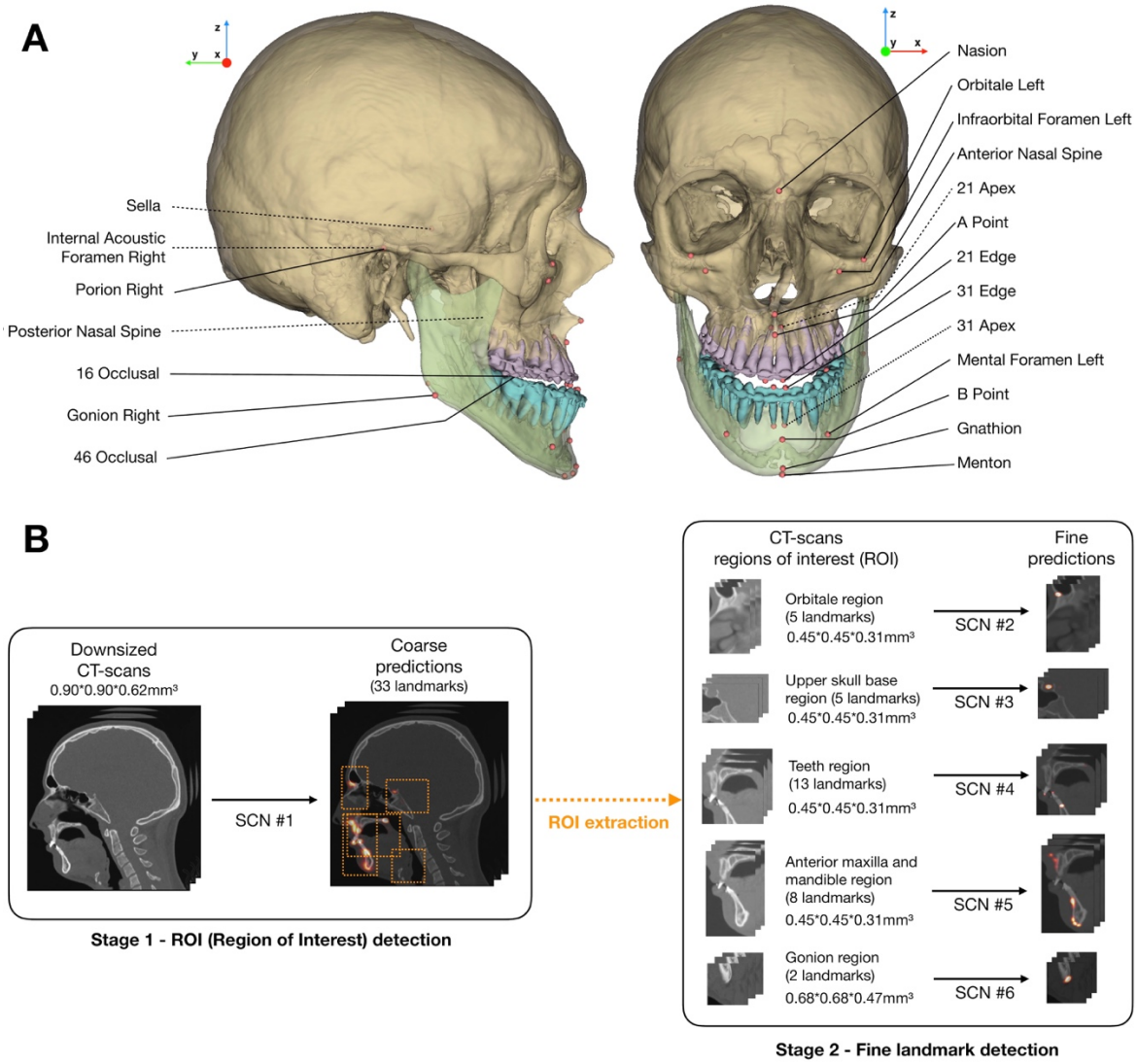


Figure 1. Landmarks and pipeline of the deep learning model. (**A**) Illustration of the set of 33 landmarks; bilateral landmarks are named once; dotted lines show landmarks localized inside the skull; (**B**) 2-stage method used for model inference. SCN, SpatialConfiguration-Net; ROI, region of interest.

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 (Appendix Table 4) 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 appendix.

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.

Results

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.

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 Appendix Tables 6, 7 and 8. 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.

Table 1. 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

Cephalometric measurements

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 2. 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)

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 (Appendix Table 9). 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 (Appendix Table 10). 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 Appendix. 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%).

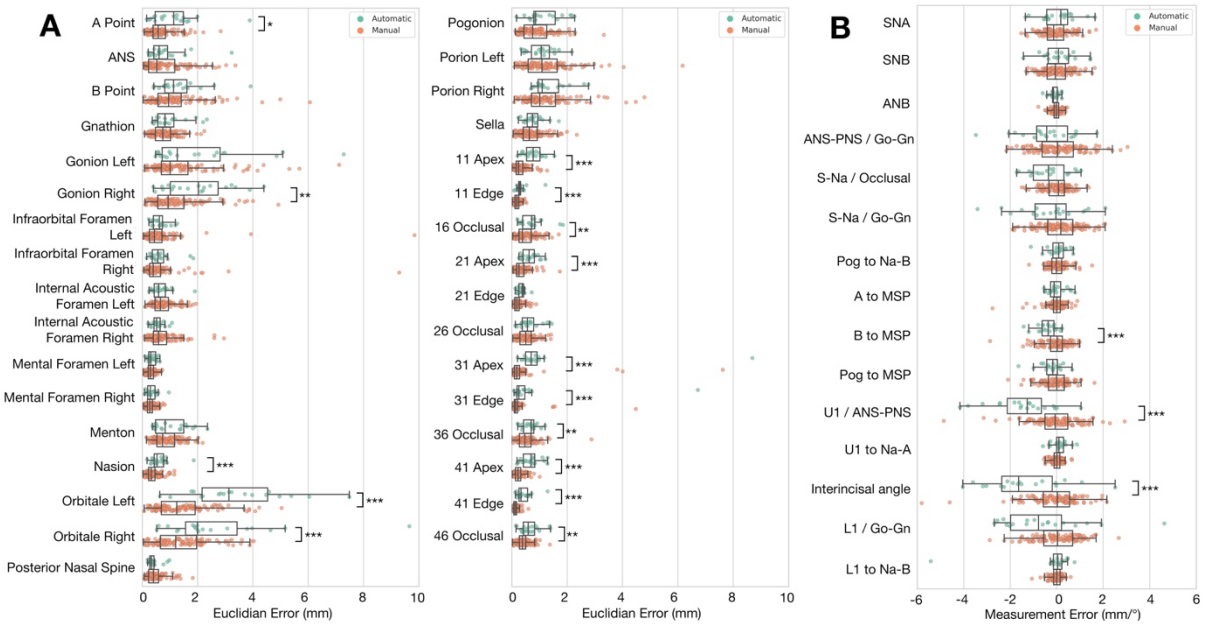


Figure 2. Localization and measurement error boxplots for automatic (green) and manual (orange) methods on 19 CT scans from the test set. (**A**) Localization errors (mm) for each landmark; (**B**) Measurement errors (mm/°) for each cephalometric variable. For each pair of results, statistically significant differences are indicated (* p<0.05; ** p<0.01; ***p<0.001).

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).

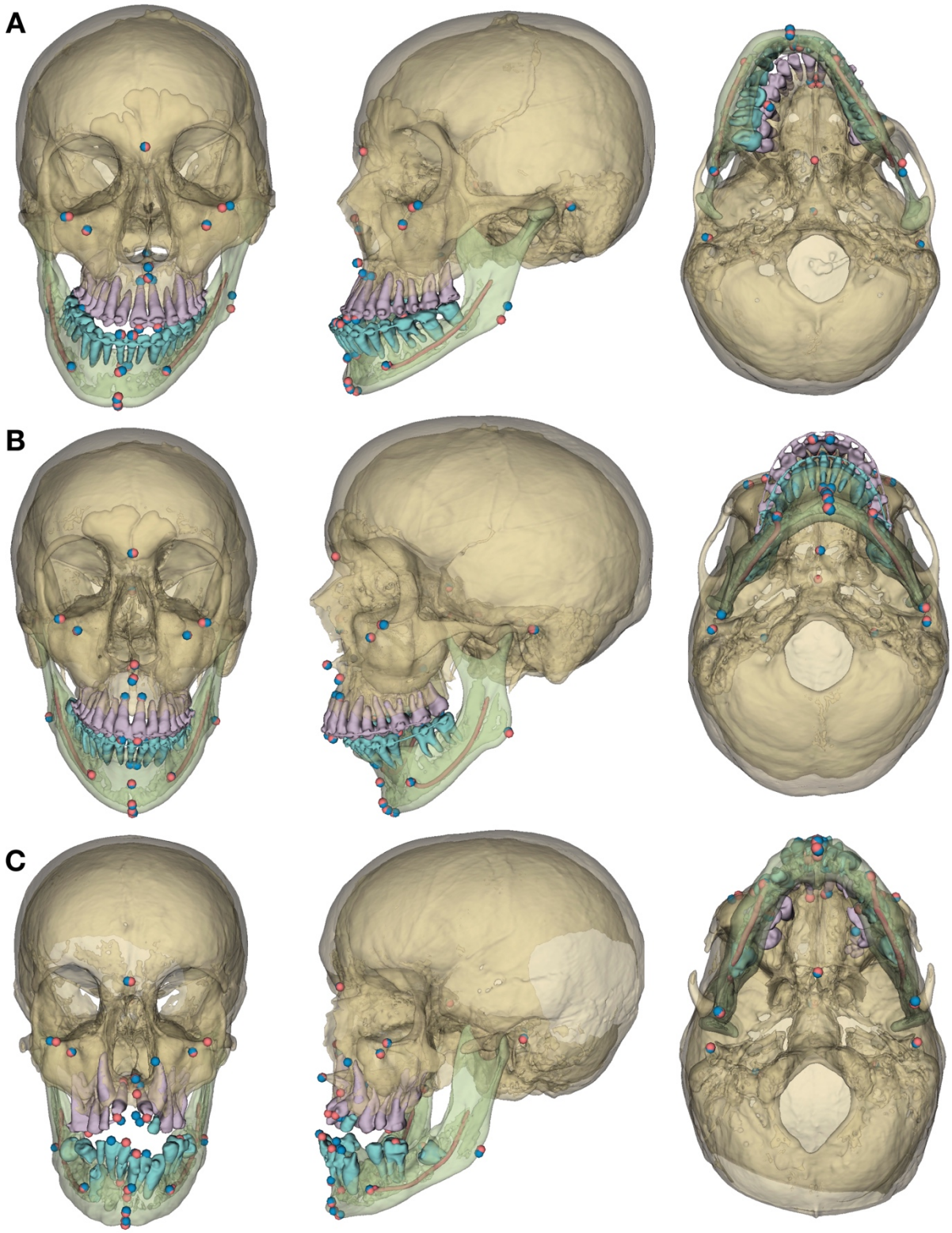


Figure 3. Frontal, ¾ left and inferior views of the 3D models, reference (red) and predicted (blue) landmarks for 3 subjects. (**A**) Prognathic and asymmetric mandible; (**B**) retrognathic mandible; (**C**) craniofacial syndrome "outlier case", the errors in the predicted A point (at the level of the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

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 (Appendix Table 7) 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.

Author Contributions

G. Dot contributed to the conception, design, data acquisition, analysis and interpretation, performed all statistical analyses, drafted and critically revised the manuscript. T. Schouman, P. Rouch and L. Gajny contributed to the conception, design and data interpretation, and critically revised the manuscript. S. Chang, F. Rafflenbeul and A. Kerbrat contributed to the data analysis and critically revised the manuscript. All authors gave final approval and agree to be accountable for all aspects of the work.

Acknowledgments

The authors would like to thank C. Payer and all the team behind SpatialConfiguration-Net for sharing their research and codes.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This study has received funding by the “Fondation des Gueules Cassées” (grant number 28–2020).

Ethical Approval

The IRB “Comité d'Ethique pour la Recherche en Imagerie Médicale” (CERIM) gave ethical approval for this research (number CRM-2001-051).

Data availability

All data produced in the present study are available upon reasonable request to the authors.

Figure and table legends

Figure 1. 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.

Figure 2. 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 (* $p<0.05$; ** $p<0.01$; ***$p<0.001$).

Figure 3. 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.

Table 1. 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.

Table 2. 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.

References

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.

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.

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.
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, 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.
Gateno J, Xia JJ, Teichgraeber JF. 2011. New 3-Dimensional Cephalometric Analysis for Orthognathic Surgery. J Oral Maxillofac Surg. 69(3):606–622.
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.
ISO 5725-2:2019. Accuracy (trueness and precision) of measurement methods and results.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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.
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, 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.
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.
Shorten C, Khoshgoftaar TM. 2019. A survey on Image Data Augmentation for Deep Learning. J Big Data. 6(1):60.
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.
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.
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.
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.
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.

Automatic 3-Dimensional Cephalometric Landmarking via Deep Learning

Gauthier Dot, Thomas Schouman, Shaole Chang, Frédéric Rafflenbeul, Adeline Kerbrat, Philippe Rouch, Laurent Gajny

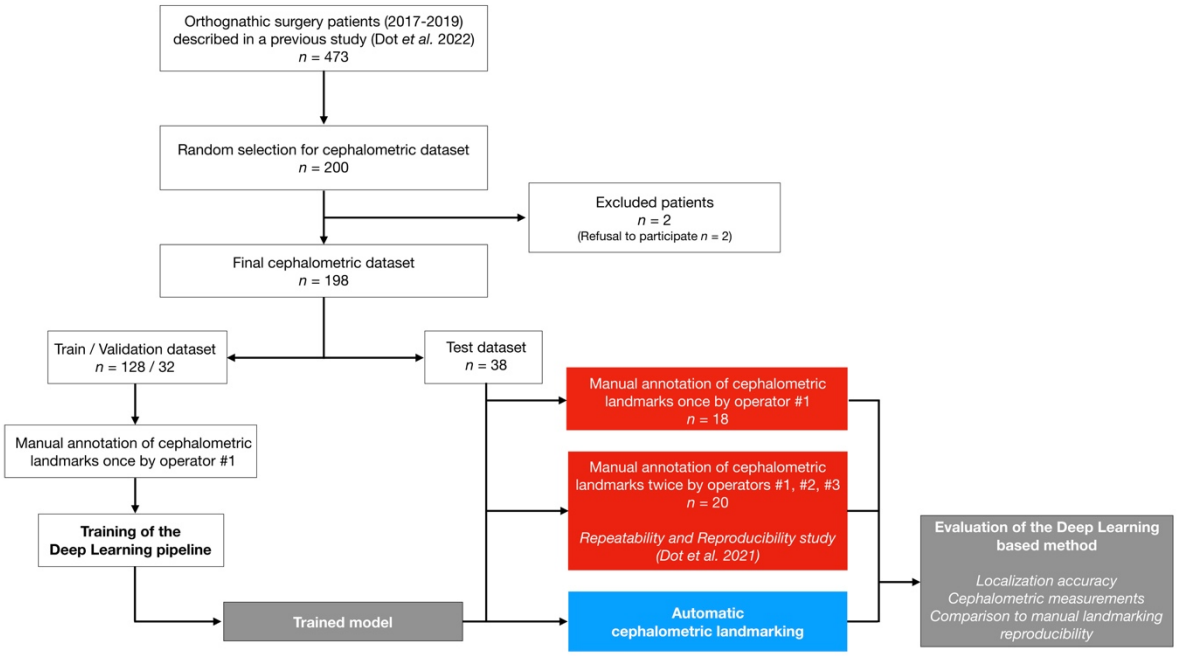
Supplemental Appendix

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.



Appendix Figure 1. 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 Appendix Table 1.

Appendix Table 1. 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 Appendix Table 2 and 3, respectively.

Appendix Table 2. 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

Appendix Table 3. 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) 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.

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

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 (Appendix Table 4) 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.

Appendix Table 4. 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 (Appendix Table 5) (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.

Appendix Table 5. 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

Appendix Table 6. 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

Appendix Table 7. 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

Appendix Table 8. 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

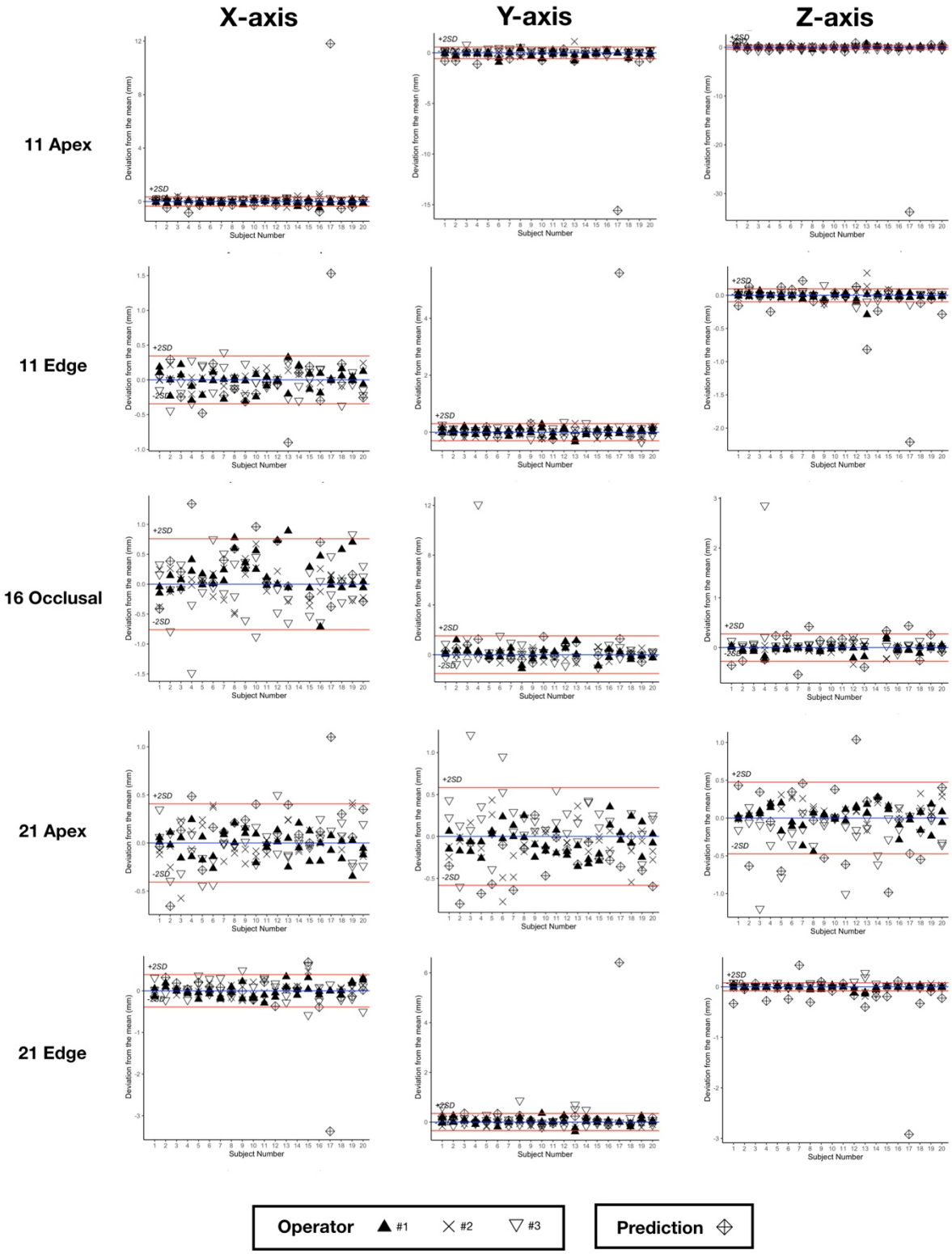
Appendix Table 9. Proportion (% (*n*)) of predicted landmarks coordinates in the -x, -y and -z directions within Bland-Altman 95% limits of agreement of manual reproducibility (95% LoA) on the hold-out test set without the outlier case (*n* = 37). L, left; R, right.

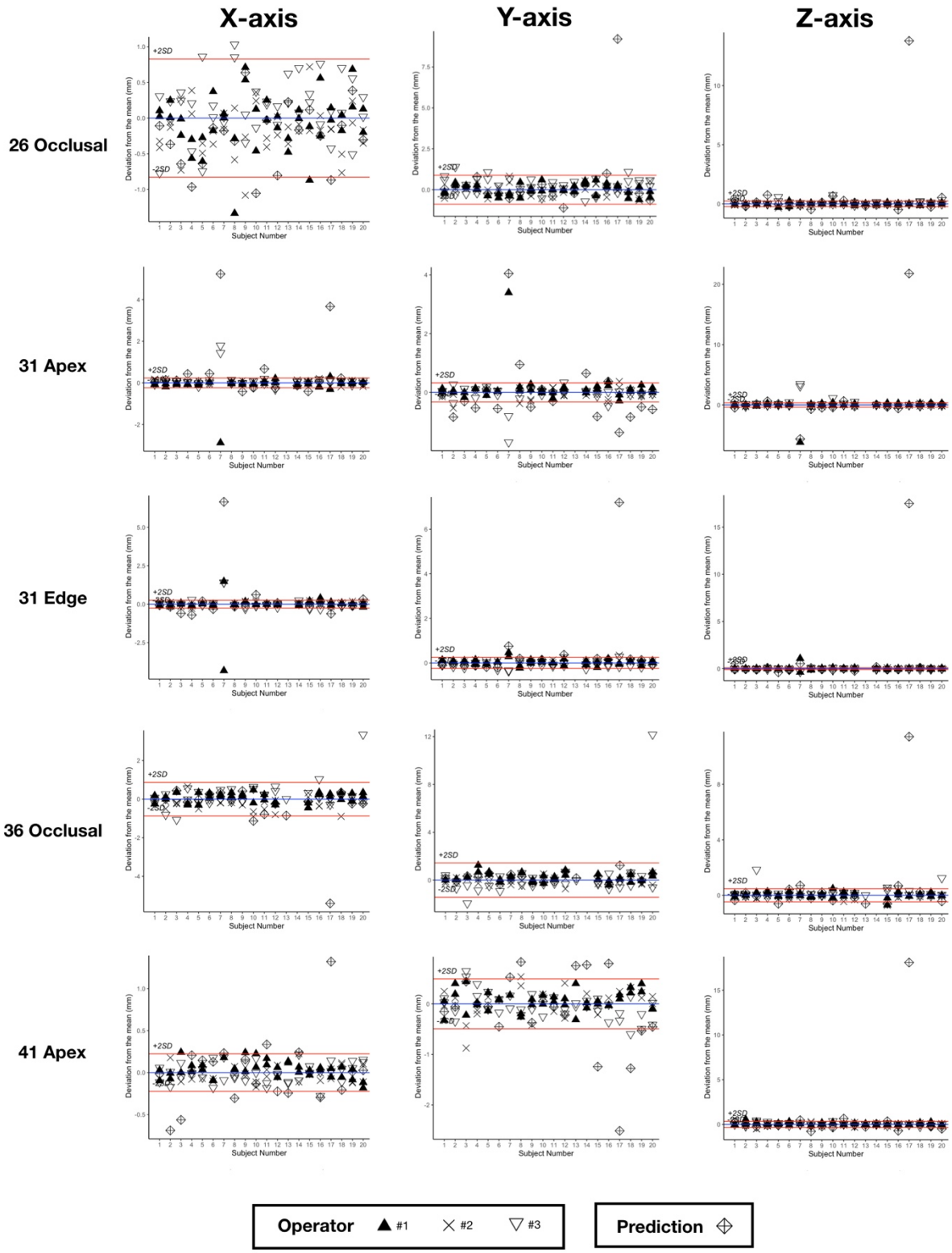
	Within 95% LoA		
	X Axis	Y Axis	Z Axis
11 Apex	67.6 (25)	67.6 (25)	62.2 (23)
11 Edge	73 (27)	81.1 (30)	40.5 (15)
16 Occlusal	81.1 (30)	91.9 (34)	56.8 (21)
21 Apex	86.5 (32)	75.7 (28)	64.9 (24)
21 Edge	73 (27)	83.8 (31)	16.2 (6)
26 Occlusal	89.2 (33)	86.5 (32)	56.8 (21)
31 Apex	59.5 (22)	27 (10)	54.1 (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)

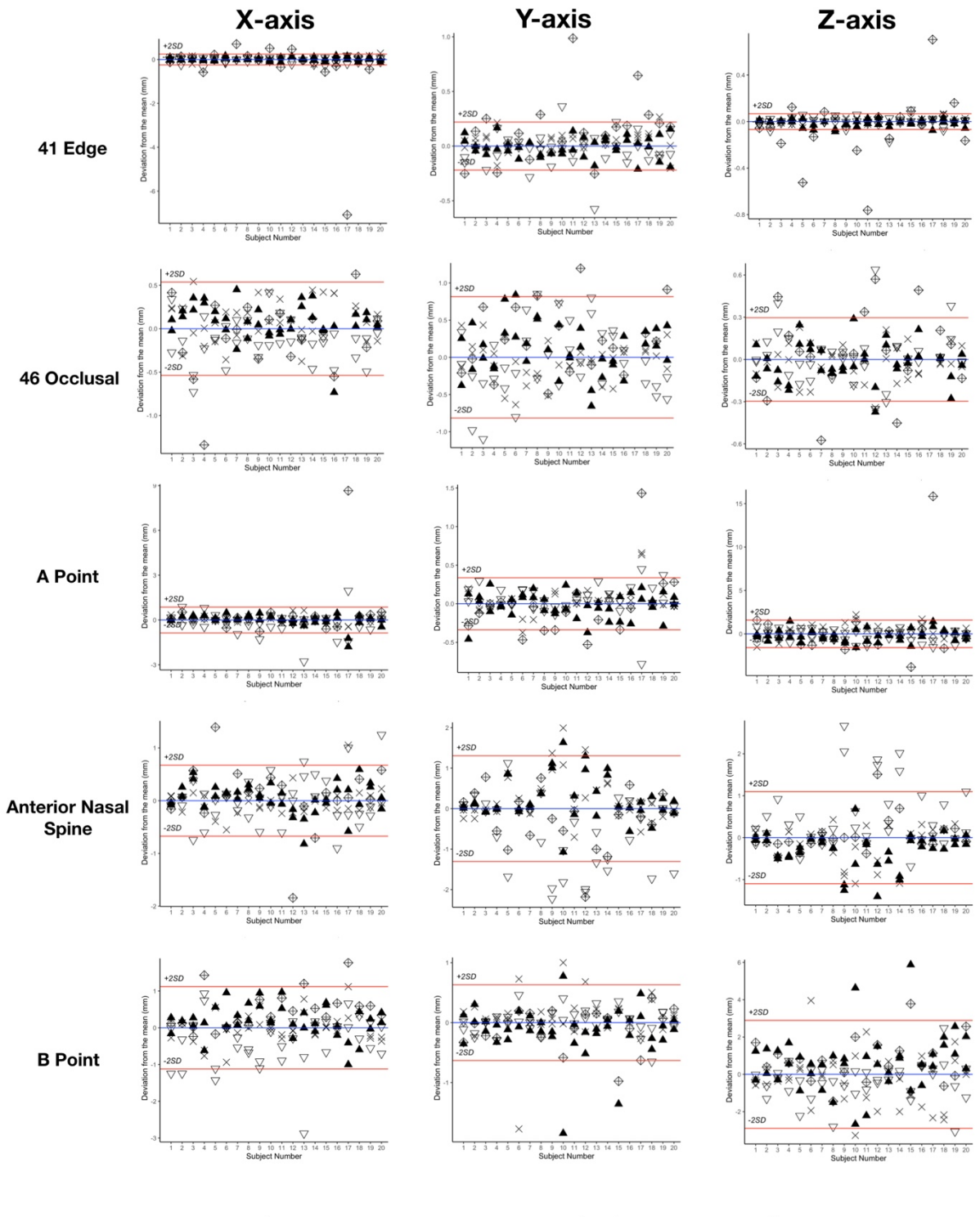
Appendix Table 10. Proportion (% (*n*)) of predicted cephalometric measurements within Bland-Altman 95% limits of agreement of manual reproducibility (95% LoA) on the hold-out test set without the outlier case (*n* = 37).

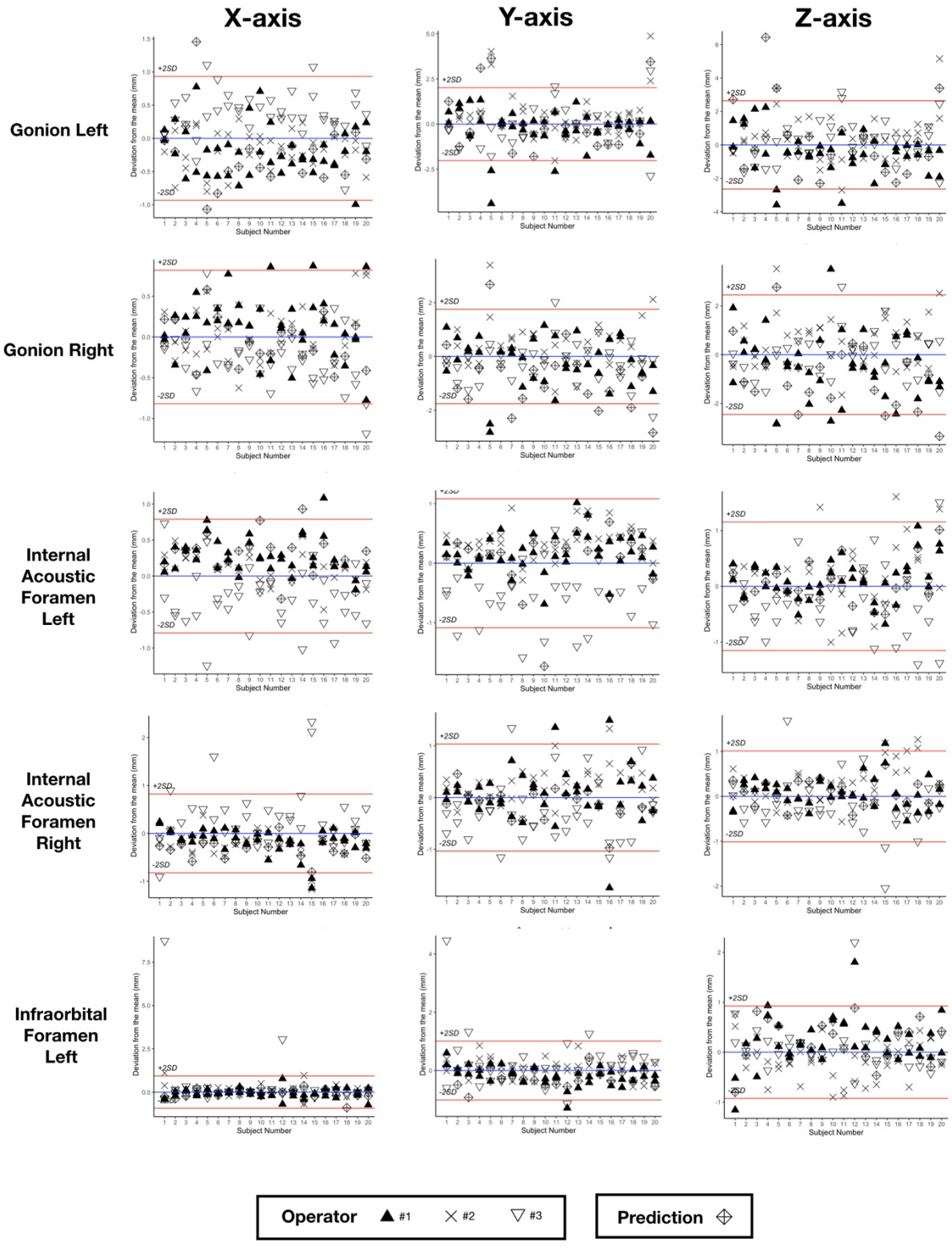
	Within 95% LoA
Skeletal	
SNA (°)	86.5 (32)
SNB (°)	91.9 (34)
ANB (°)	91.9 (34)
ANS-PNS / Go-Gn (°)	94.6 (35)
S-Na / Go-Gn (°)	81.1 (30)
Pog to Na-B (mm)	94.6 (35)
A to MSP (mm)	100 (37)
B to MSP (mm)	91.9 (34)
Pog to MSP (mm)	94.6 (35)
Dentoalveolar	
S-Na / Occlusal plane (°)	62.9 (22)
U1 / ANS-PNS (°)	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)

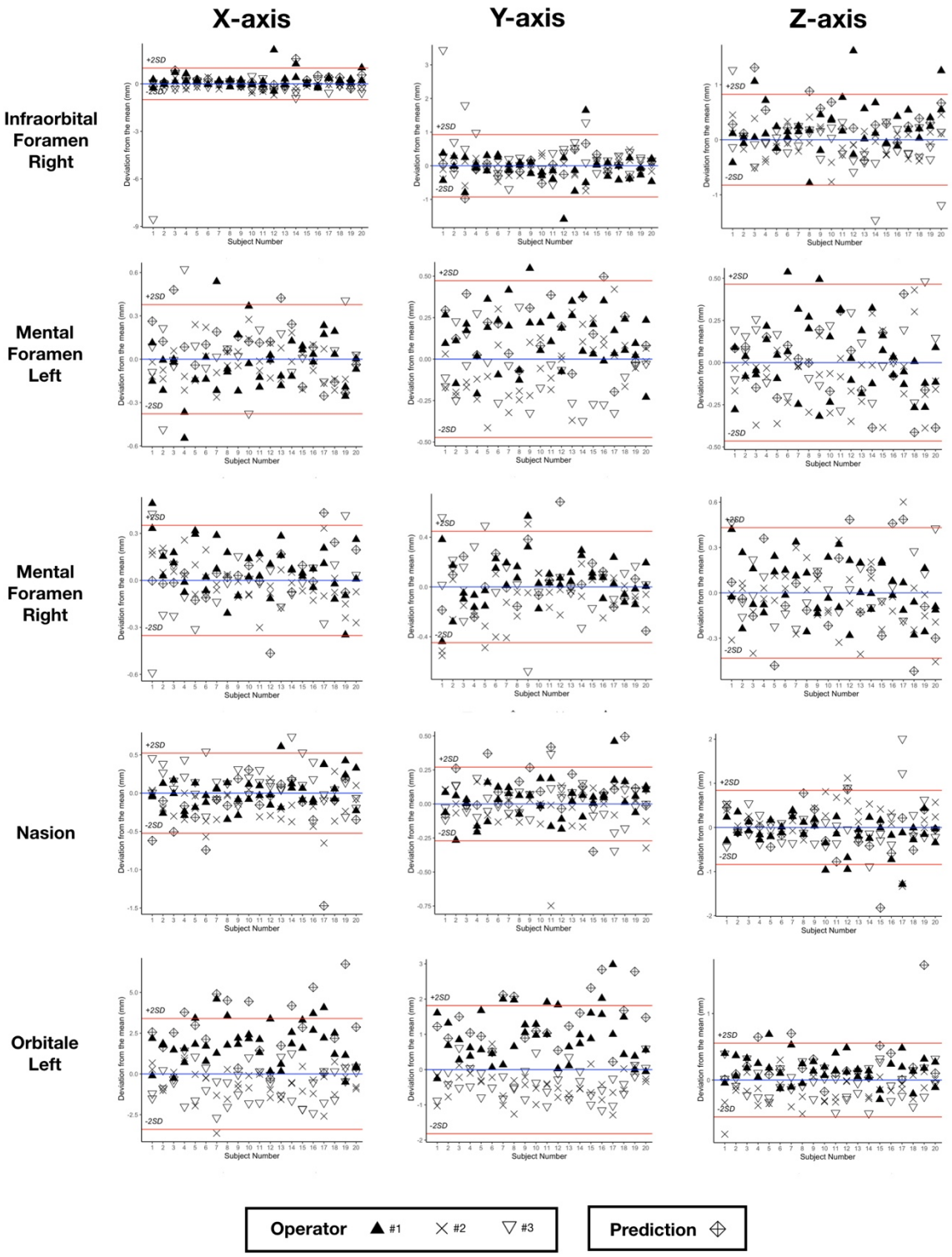
Bland-Altman plots of landmarking localization. For the 33 landmarks, the following plots show the deviations from the mean (blue line) of the 6 manual repetitions and the predictions for the 20 subjects included in the R&R study. Please note that the scales differ. Subject number 17 is the outlier case.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

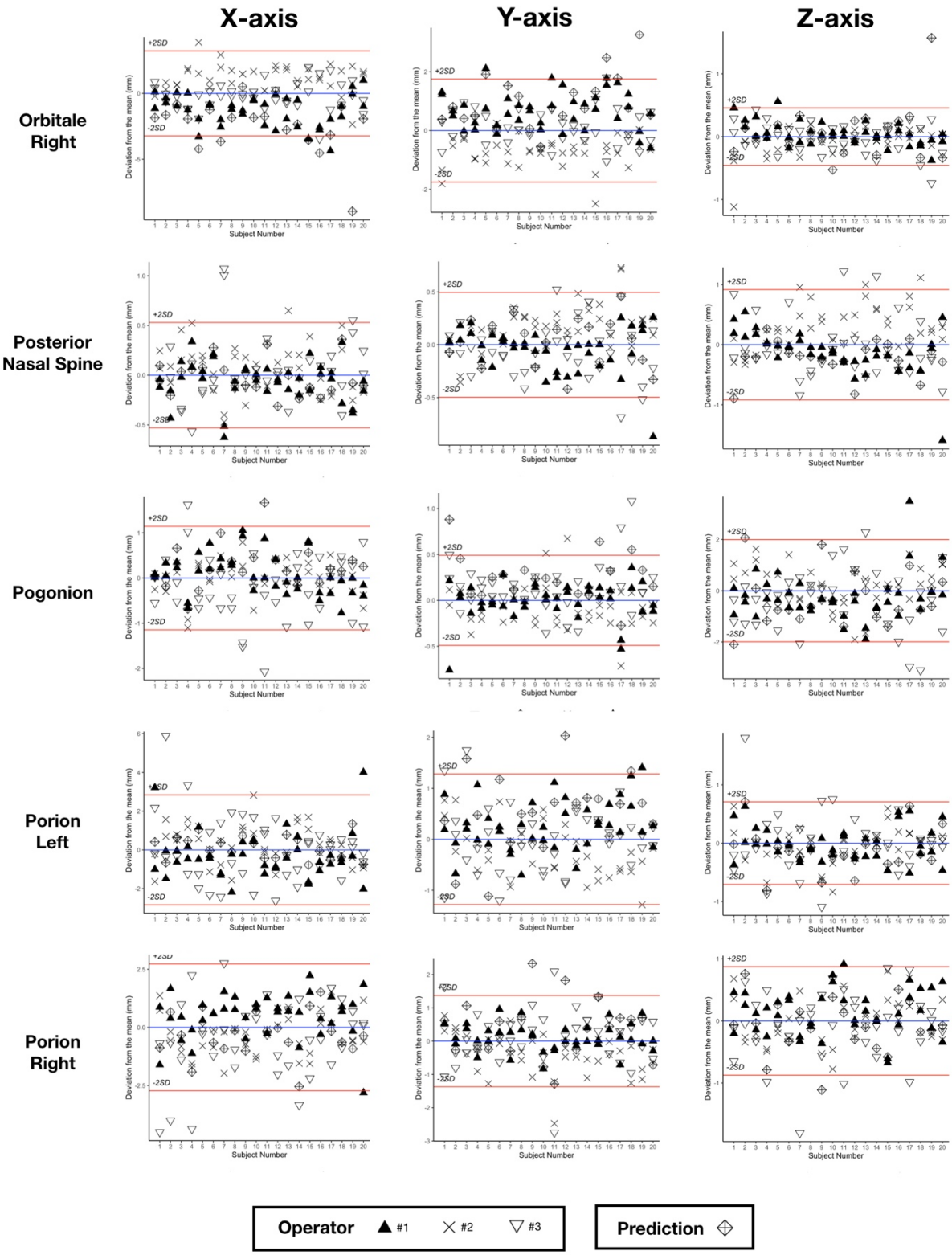


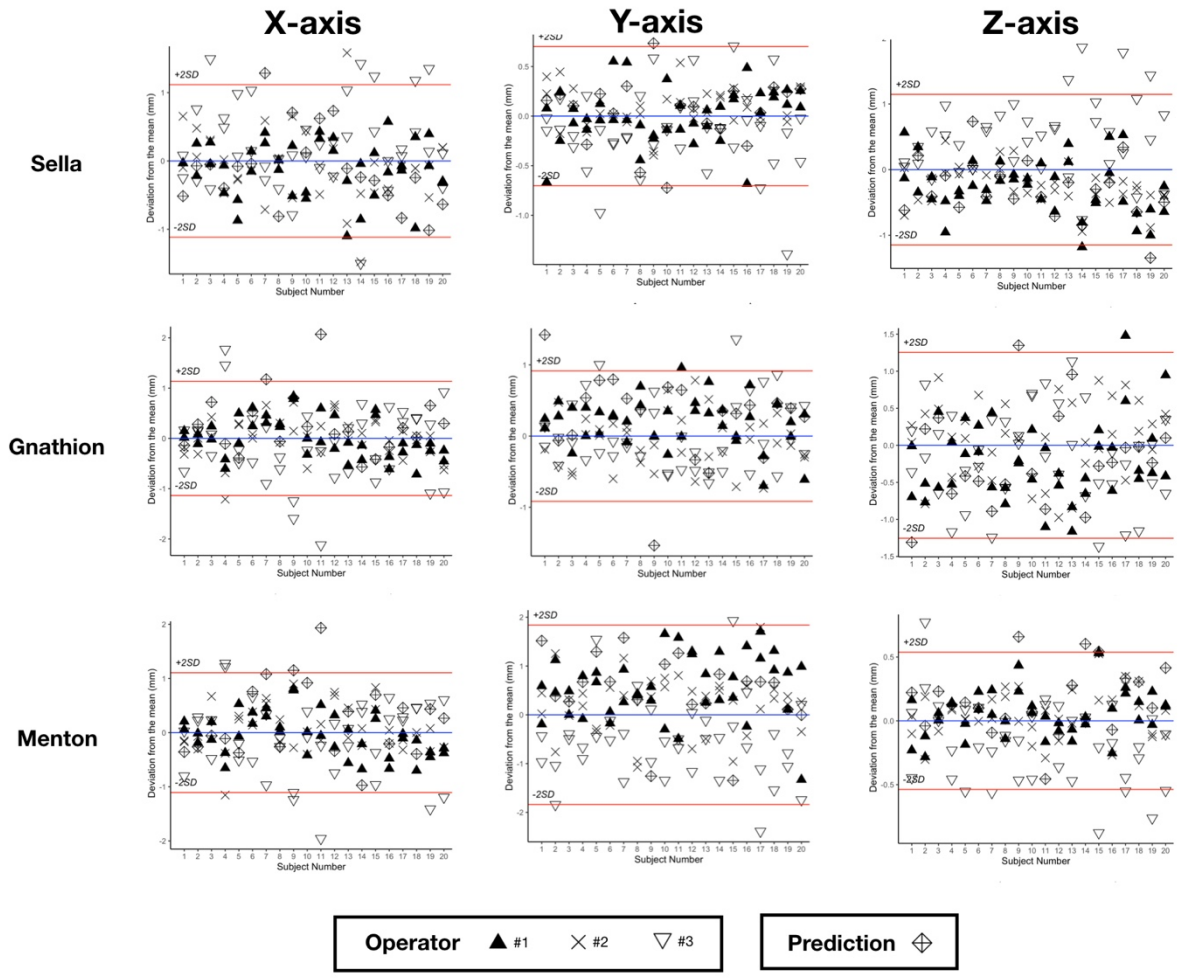

X-axis
Y-axis
Z-axis
26 Occlusal
31 Apex
31 Edge
36 Occlusal
41 Apex
Deviation from the mean (mm)
Subject Number
+2SD
-2SD
Operator
#1
#2
#3
Prediction


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

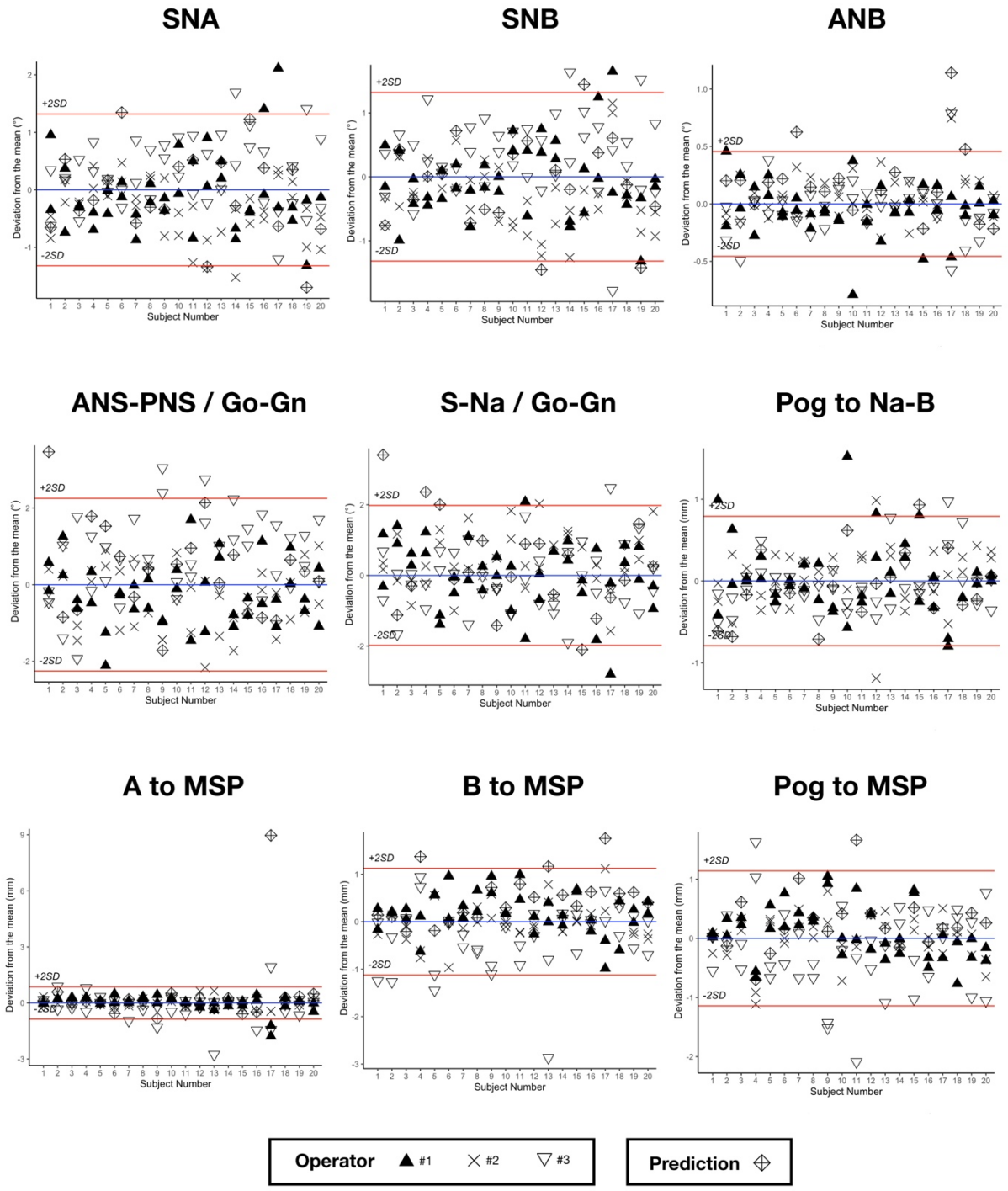

X-axis
Y-axis
Z-axis
Gonion Left
Gonion Right
Internal Acoustic Foramen Left
Internal Acoustic Foramen Right
Infraorbital Foramen Left
+2SD
-2SD
Deviation from the mean (mm)
Subject Number
Operator
#1
#2
#3
Prediction

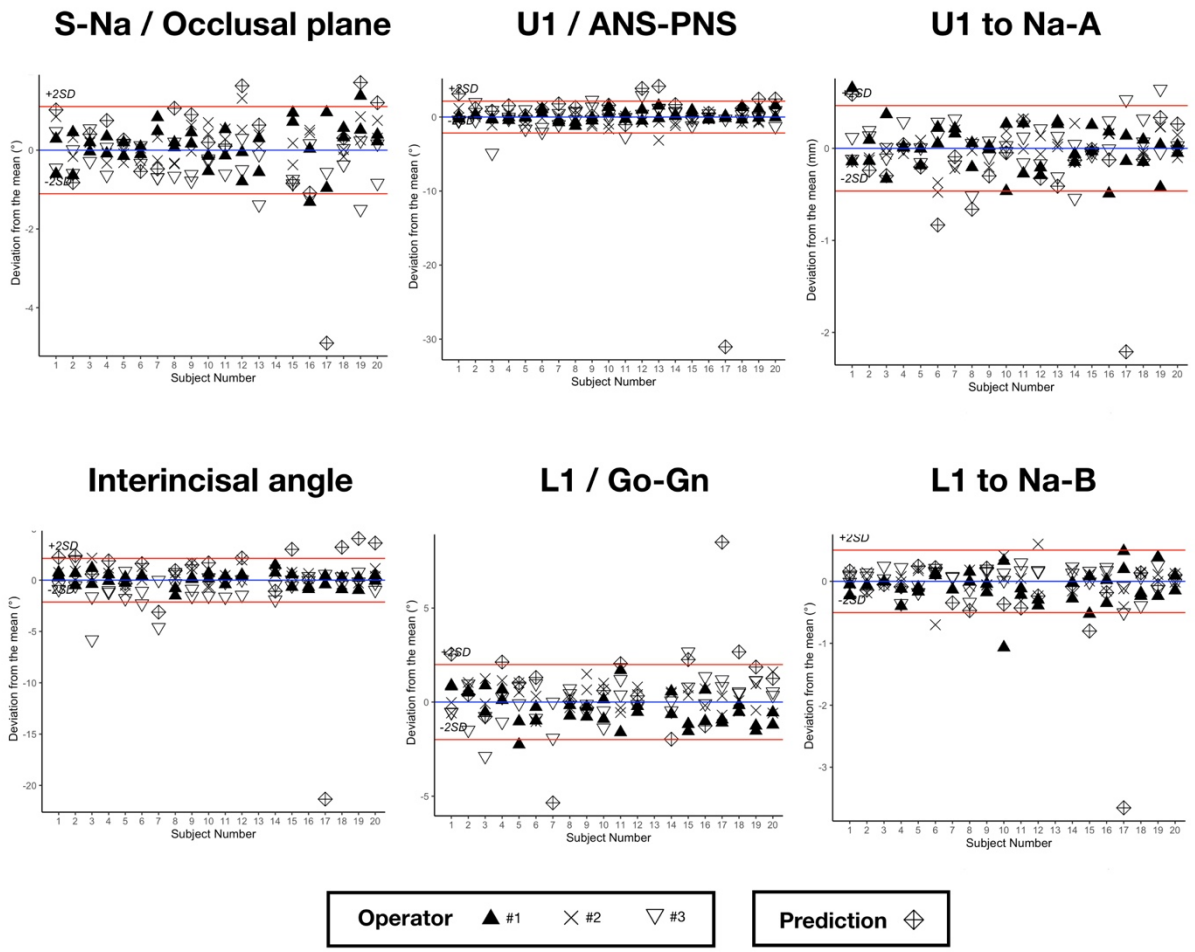

X-axis
Y-axis
Z-axis
Infraorbital Foramen Right
Mental Foramen Left
Mental Foramen Right
Nasion
Orbitale Left
Deviation from the mean (mm)
Subject Number
+2SD
-2SD
Operator
#1
#2
#3
Prediction


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


X-axis
Y-axis
Z-axis
Sella
Gnathion
Menton
Deviation from the mean (mm)
Subject Number
+2SD
-2SD
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.





References

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, 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.

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.