Original Article

Beyond Benchmarks: Towards Robust Artificial Intelligence Bone Segmentation in Socio-Technical Systems

Kunpeng Xie[1,2], Lennart Johannes Gruber[3], Martin Crampen[2,1], Yao Li[1,2], André Ferreira[1,2,4,10], Elias Tappeiner[7], Maxime Gillot[31,32], Jan Schepers[5], Jiangchang Xu[6], Tobias Pankert[8, 1], Michel Beyer[9,37], Negar Shahamiri[10], Reinier ten Brink[11, 34], Gauthier Dot[12], Charlotte Weschke[13], Niels van Nistelrooij[14,18,33], Pieter-Jan Verhelst[15,35], Yan Guo[6,1], Zhibin Xu[1], Jonas Bienzeisler[2], Ashkan Rashad[1], Tabea Flügge[18.33], Ross Cotton[17], Shankeeth Vinayahalingam[14], Robert Ilesan[27], Stefan Raith[8], Dennis Madsen[19], Constantin Seibold[10], Tong Xi[14], Stefaan Bergé[14], Sven Nebelung[25], Oldřich Kodym[20], Osku Sundqvist[24], Florian Thieringer[9,37], Hans Lamecker[13], Antoine Coppens[16], Thomas Potrusil[23,36], Joep Kraeima[11, 34], Max Witjes[11], Guomin Wu[22], Xiaojun Chen[6], Adriaan Lambrechts[5], Lucia H Soares Cevidanes[31], Stefan Zachow[21,18], Alexander Hermans[26,25], Daniel Truhn[25], Victor Alves[4], Jan Egger[10,29,30], Rainer Röhrig[2], Frank Hölzle[1], Behrus Puladi[1,2]✉

1.Department of Oral and Maxillofacial Surgery, University Hospital RWTH Aachen, 52074 Aachen, Germany; 2.Institute of Medical Informatics, University Hospital RWTH Aachen, 52074 Aachen, Germany; 3.Department of Oral and Maxillofacial Surgery, University Medical Center Goettingen, 37075 Goettingen, Germany; 4.Center Algoritmi/LASI, University of Minho, 4710-057 Braga, Portugal ; 5.Materialise NV, 3001 Leuven, Belgium; 6.School of Mechanical Engineering, Shanghai Jiao Tong University, 200240 Shanghai, China; 7.UMIT TIROL - Private University For Health Sciences and Health Technology, 6060 Hall in Tyrol, Austria; 8.Inzipio GmbH, 52070 Aachen, Germany; 9.Department of Oral and Cranio-Maxillofacial Surgery, University Hospital Basel, 4031 Basel, Switzerland; 10.Institute of Artificial Intelligence in Medicine (IKIM), University Hospital Essen, 45131 Essen, Germany; 11.Department of Oral & Maxillofacial Surgery, University Medical Center Groningen, 9713 GZ Groningen, The Netherlands; 12.Universite Paris Cité, UFR Odontologie, F-75006 Paris, France; 13.1000shapes GmbH, 12247 Berlin, Germany; 14.Department of Oral and Maxillofacial Surgery, Radboud University Medical Center, 6525 GA Nijmegen, The Netherlands; 15.Department of Oral and Maxillofacial Surgery, University Hospitals Leuven, 3000 Leuven, Belgium; 16.Relu BV, 3001 Leuven, Belgium; 17. Synopsys Northern Europe Ltd., Exeter EX4 3PL, United Kingdom; 18.Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt Universität zu Berlin, Department of Oral and Maxillofacial Surgery, 12203 Berlin, Germany; 19.University of Zürich, CH-8006 Zürich, Switzerland; 20.TESCAN 3DIM, 61700 Brno, Czech Republic; 21.Department of Visual- and Data-centric Computing, Zuse Institute, 14195 Berlin, Germany; 22. Craniomaxillofacial Plastic and Cosmetic Center, Hospital of Stomatology, Jilin University, 130021 Changchun, China; 23.CADS GmbH, 4320 Perg, Austria; 24.Planmeca Oy, FIN-00880 Helsinki, Finland; 25.Department of Interventional and Diagnostic Radiology, RWTH Aachen University, 52074 Aachen, Germany; 26.Visual Computing Institute, Computer Science and Natural Sciences, RWTH Aachen University, 52074 Aachen, Germany; 27.Department of Oral and Maxillofacial Surgery, Lucerne Cantonal Hospital, 6000 Lucerne, Switzerland; 28.Center for Virtual and Extended Reality in Medicine (ZvRM), Essen University Hospital (AöR), 45147 Essen, Germany; 29.Cancer Research Center Cologne Essen (CCCE), University Medicine Essen (AöR), 45147 Essen, Germany; 30.University of Duisburg-Essen, Faculty of Computer Science, 45127 Essen, Germany; 31.University of Michigan, Ann Arbor, 48109-1079 Michigan, United States; 32.CPE Lyon, 69100 Villeurbanne, Lyon, France; 33.Einstein Center Digital Future (ECDF), 10117 Berlin, Germany; 34.3D Lab University Medical Center Groningen, University of Groningen, 9713 GZ Groningen, The Netherlands; 35.OMFS-IMPATH Research Group, Department of Imaging and Pathology, Faculty of Medicine, KU Leuven, 3000 Leuven, Belgium; 36.KLS Martin Group, 78532 Tuttlingen, Germany; 37.Medical Additive Manufacturing Research Group (Swiss MAM), Department of Biomedical Engineering, University of Basel, 4123 Allschwil, Switzerland;

✉**Corresponding author:**
Dr. med. Dr. med. dent. Behrus Puladi
OCRID ID: 0000-0001-5909-6105
E-mail: bpuladi@ukaachen.de
Tel.: +49-241-80-38389
Fax: +49-241-80-82430
Department of Oral and Maxillofacial Surgery & Institute of Medical Informatics,
University Hospital RWTH Aachen, Pauwelsstraße 30, 52074 Aachen, Germany

Key words: Artificial Intelligence; Image Processing, Computer-Assisted; Imaging, Three-Dimensional; Tomography, X-Ray Computed; Practice Guideline; Mandible

Abstract:

Despite the advances in automated medical image segmentation, AI models still underperform in various clinical settings, challenging real-world integration. In this multicenter evaluation, we analyzed 20 state-of-the-art mandibular segmentation models across 19,218 segmentations of 1,000 clinically resampled CT/CBCT scans. We show that segmentation accuracy varies by up to 25% depending on socio-technical factors such as voxel size, bone orientation, and patient conditions such as osteosynthesis or pathology. Higher sharpness, isotropic smaller voxels, and neutral orientation significantly improved results, while metallic osteosynthesis and anatomical complexity led to significant degradation. Our findings challenge the common view of AI models as "plug-and-play" tools and suggest evidence-based optimization recommendations for both clinicians and developers. This will in turn boost the integration of AI segmentation tools in routine healthcare.

Introduction

With the ongoing digital transformation of healthcare, segmentation-based acquisition of anatomical and pathological structures has become an essential step in both clinical practice and research. Applications scenarios span over a wide field including diagnostic, image-guided radiotherapy and virtual surgical planning[1–3]. However, manual segmentation is still labor intensive and time-consuming. To address this issue, a large number of automatic segmentation methods for different structures have emerged in the last decades, and among them artificial intelligence (AI) models utilizing deep learning methods are the most promising ones[4–6]. In the segmentation of mandible for example, AI models have progressed beyond research settings and have begun to translate to clinical use as certified medical software in clinical practice[7–10]. However, despite decades of algorithmic advancements, there remains no standardized clinical integration protocol for AI segmentation models, leaving clinical integration a major challenge[5,11,12].

This may be due to the technocentric paradigm that has been in place for decades of comparing and developing algorithms in different challenges to push the limits of performance and ultimately surpass human capabilities[13]. While this technocentric perspective has brought us powerful models and refreshed leaderboards, it often overlooks the complex socio-technical systems in which AI models are applied. In terms of clinicians, recent work shows that their adoption of AI generated results hinge on transparency, robustness, and real-world applicability—not benchmark metrics alone[14,15]. Additionally, in real-world situations, medical imaging data is often acquired prospectively based on specific clinical needs, including a wide range of possible imaging protocols as well as different patient factors. In this respect, a shift in perspective from a techno-centric preoccupation to a socio-technical perspective[16], which explicitly considers clinical contexts such as diverse imaging protocols, patient demographics, and practical workflow integration, would be highly beneficial in facilitating the effective translation of AI segmentation models into clinical routines and research settings.

Consequently, we need to understand how socio-technical factors affect the performance of AI segmentation models in general. A previous study found that factors such as the imaging modalities (e.g., CT and CBCT), scanning devices, and the reconstruction protocols (e.g., voxel size, thickness, convolutional kernels) all may impact segmentation outcomes[17]. While some studies have begun to explore these factors, previous studies have either focused on limited factors or used only a single AI model, leaving a comprehensive understanding of these interactions largely unveiled[18,19].

To address this issue, instead of simply comparing models' performance, we evaluated the impact of socio-technical factors on the overall performance of multiple AI models in this study. For this purpose, we chose the mandible, which is morphologically complex and a representative in bone segmentation, as the segmentation target and created a benchmark dataset that balanced both patient and imaging features. Notably, our study recruited the largest number of AI models for mandible segmentation evaluated to date. By systematically resampling the original data, we could experimentally control the impact of different factors as they would be controllable during medical image acquisition. We then evaluated the segmentation results to explore the general impact of imaging, patient, and anatomical region factors on the model performance. Based on the results, we further suggest best practice recommendations for clinicians in applying AI segmentation models. In addition, we put forward requirements for AI developers, who are expected to create next-generation models that are informed by the clinical challenges encountered in AI models. Our study provides a reliable evidence base for future clinical integration guidelines of AI segmentation models, helping bridge the gap between technical performance and practical deployment.

Methods

In this multicenter study we evaluated state-of-the-art AI models from 20 different centers and companies around the world (Table 1). The study protocol was registered prospectively in the German clinical trial registry under registration ID DRKS00032736. All technical details can be found in this study protocol. The ethics application of the study was approved by the ethics committee at RWTH Aachen University (No.23-272). No informed consent was needed due to the use of anonymized retrospective patient data.

Dataset Preparation

To build a balanced benchmark dataset in terms of patient-related features, we selected 50 CT and 50 CBCT scans from 100 patients from a single center. In terms of patient characteristics, the sex ratio is 1:1 and the average age of patients was 48.47 years (range 19 – 91 years) (Supplementary Table 1). All selected scans were de-identified by cropping out the region above the inferior border of the orbital rim. Cases were excluded if cropping was not possible without affecting the condyle region. We systematically resampled the original 100 selected cases to create an additional 900 volumes, for a total of 1,000 volumes. This method, instead of selecting 1,000 cases directly, gave us full control over the voxel size, slice thickness, sharpness, noise and rotation of the mandible.

To obtain a balanced dataset, the features of the original CT/CBCT volumes were profiled prior to resampling. These features were quantified and measured in five aspects: a) voxel size (XY); b) slice thickness; c) sharpness; d) noise; e) rotation of the mandible. Where a) and b) were extracted from DICOM tags, c) was quantified via a Sobel-based edge intensity, and d) was derived from the standard deviation of the median-filter difference. Mandible rotations e) were calculated using bone landmarks. Based on the measurements, we chose five types of resampling methods namely: a) increase the slice thickness; b) expand the voxel size (XY); c) sharpening / smoothing; d) Gaussian-noise / denoise; e) rotation in axial, coronal and sagittal plane. A set of factors were tested and used in resampling these features respectively (Supplementary Table 4). By adjusting these factors, we have managed to approximate the distributions of features on the resampled dataset to the reference distribution from public datasets [20–22] or normal distribution. A total of 3,727,360 resampling combinations of imaging features were generated, from which 900 were randomly selected and resampled volumes were generated accordingly (Supplementary Figure 1). These down-sampled volumes, together with the initial 100 scans, resulted in a balanced final dataset of 1,000 volumes. The final distribution of patient and imaging features can be found in Supplementary Figure 1.

Ground Truths

Mandible segmentations of the original scans were performed by two surgeons experienced in segmentation (KX and LG) independently in different software, KX in Mimics (Version 21.0) and LG in 3D Slicer (Version 5.6.2). The quality of segmentations was checked and approved by a third surgeon (BP). The principle of the segmentation was to preserve the anatomical bone structure of the mandible. In this case, all teeth, including dental implants, crowns and bridges, were segmented along with the mandible. Osteosynthesis materials (e.g. reconstruction plates, fixation screws/plates) were excluded in the segmentation, except for the part inside the mandible. The cancellous bone and the mandibular nerve canal were filled in so that the final segmentation result is free of internal cavities. Since resampling is not changing the anatomy of the bone, we applied the same resampling protocol in voxel scaling and rotation to the original ground truths to obtain corresponding segmentation results for the resampled 900 cases.

Model Recruitment

The segmentation models included in this study need to meet the following criteria: a) deep learning based fully automatic segmentation tool; b) developed within the last five years; c) the output of the model is the mesh model or label map of the whole mandible; d) already trained and ready to use. Based on the literature study of a systematic review, we listed a group of models available in publications and searched further in online databases for other models published after the systematic review[5]. We contacted 35 corresponding authors and ten of them agreed to participate in the study. In addition, ten companies that offer mandible segmentation tools as a service were contacted. Eight of them joined our study. Furthermore, we have searched public repositories for available models and applied two trained models. With a data transfer agreement (DTA), the final dataset was shared with the collaborators, and segmentation results were returned to RWTH Aachen for evaluation. If a DTA was not feasible or the model was publicly available, inference was conducted locally at RWTH Aachen University.

Evaluation

To further evaluate the segmentation quality in different anatomical regions of the mandible, we delineated nine ROIs by K.X and controlled by B.P.: condyle L/R, inferior alveolar nerve (IAN) entrance L/R, IAN exits L/R, dentition, inferior border. The last ROI, mandible body, was defined as the rest of mandible excluding the ROIs. All of the above ROIs were created based on reference points manually labelled on the volume by KX. Segmentation results were compared to both manual ground truths and the mean value was taken as the final result. We chose four metrics for evaluation: DSC, NSD, HD95, and MASD, and all metrics were calculated using the python package from Nikolov et al.[23,24] on the whole mandible and on all ROIs respectively. No evaluations in the dentition region were conducted if the AI model cannot segment the teeth. All evaluations were conducted anonymously to secure the interests of all researchers and companies.

Statistical analysis

The statistical analysis was conducted with the R programming language (Version 4.4.2). For descriptive statistics on data with non-normal distribution, we applied the non-parametric Mann-Whitney U test to evaluate statistical significance, followed by a bootstrap procedure with 5000 replicates to obtain the 95% CI for the median difference. Factors listed in the above section were set as fixed effect in the LMM while the difference of the AI models was considered as random effect. We checked the collinearity of selected fixed effects and found that sharpness and noise were highly corelated with a Variable Inflation Factor (VIF) of 10.943. In this case, noise was removed from the list of factors. We scaled the factors and tested multiple combinations of settings and selected one optimal LMM for each ROI and the whole mandible on each metric. LMM results on DSC are displayed in Figure 7. LMMs on other metrics and details of fitted models are described in Supplementary Figure 2, 3 and 4. We performed further analyses of the models described above to establish the evidence base for our recommendations.

Results

Recruited AI Models and overall segmentation results

A total of 20 commercial and research AI models for mandible segmentation from different countries across the world were recruited in this study, with the workflow shown in Figure 1. All models were developed over the last 5 years and listed in Table 1. Due to privacy reasons of the participating companies, evaluations of these models were anonymized. The evaluation was performed on ground truths of two investigators with an interrater correlation of 95.7% in Dice Similarity Coefficient (DSC, i.e. overlap measurement). From the 1000 volumes to be segmented, on average 942 volumes were successfully segmented and 19,218 segmentations were evaluated. The model designations are listed in descending order according to the number of volumes with DSC greater than 90% in their segmentation results (Fig. 2a). Only one model (S) was unable to segment any CBCT volume.

Table 2 presents the overall performance of the models, including the CT and CBCT subsets. The metrics used were: DSC as primary metric, Normalized Surface Dice (NSD, i.e. boundary agreement), 95 percentile Hausdorff Distance (HD95, i.e. worst-case boundary error), and Mean Average Surface Distance (MASD, i.e. average boundary deviation)[25]. The mean values of DSC and NSD for all models are both 81.7%. While the mean values of HD95 and MASD are 14.89 mm and 2.73 mm, respectively. Model *A* demonstrates the best performance across almost all metrics. We explored the effect of the type of training data on the segmentation results (Fig. 2b, c). It is interesting to note that the models trained with only CBCT data show better results than the models trained with only CT data (Mann-Whitney U test, $p < 0.001$), and the median difference was estimated as 5.10% with a 95% bootstrap confidence interval (CI) of [4.71%, 5.51%]. Yet the difference is not significant between CBCT and combination of both data modalities (Mann-Whitney U test, $p = 0.733$). Commercial models demonstrate better performance compared to research models (Mann-Whitney U test, $p < 0.001$), with a median difference of 1.03% [95% CI: 0.75%, 1.34%]. Regarding the amount of training data, the models trained on a moderate number of scans (150–300 cases) exhibited the optimal segmentation performance among all groups ($p < 0.001$, Kruskal-Wallis test; Fig. 2d,e). The median DSC difference between the medium and low groups was 2.70% [95% CI: 2.39%, 2.97%], and between the medium and high groups was 2.87% [95% CI: 2.55%, 3.16%].

Imaging factors

Figure 3 shows the effect of imaging factors on segmentation performance of AI models. Higher sharpness level generally leads to better segmentation results (Fig. 3c). Further analysis in Linear Mixed-effect Models (LMMs) shows a 0.50% increase in DSC per 500 Hounsfield Unit (HU)/mm increase in sharpness (LMM, β = 0.001%, $p < 0.001$). However, the DSC improvements reached a plateau beyond a certain sharpness level (approximately 5000 HU/mm). This pattern was also observed regarding noise, where a moderate noise level led to the best segmentation performance. Larger voxel sizes in the XY plane significantly reduced segmentation performance, with a 0.16% decrease in DSC for every 0.1 mm increase in in-plane voxel size (LMM, β = -1.62%, $p < 0.001$). Increasing slice thickness also had a negative impact, with DSC declining by 0.10% for every 0.1 mm increase in slice thickness (LMM, β = -0.955%, $p < 0.001$). Rotation of the mandible in all three planes resulted in negative effects on segmentation performance, with axial and sagittal rotations reducing DSC by 0.51% and 0.69% per 5-degree increase, respectively (LMM, β_{axial} = -0.102%, $\beta_{sagittal}$ = -0.138%, $p < 0.001$), while coronal rotation had no significance ($p = 0.520$).

In univariable descriptive statistics, the AI models showed better performance on CBCT data than that of CT data (Mann-Whitney U test, $p < 0.001$; Fig. 3a), with a median DSC difference of 3.20% [95% CI: 2.96%, 3.45%]. For the use of different CBCT devices, no significant difference was found (Mann-Whitney U test, $p = 0.198$; Fig. 3b). Yet a marginal decline of 1.43% in median DSC [95% CI: 1.02%, 1.78%] is found in CT device C (Mann-Whitney U test, $p < 0.001$). However, in multivariable analysis the segmentation performance of the AI model on CT data is improved by 4.13% compared to CBCT data, (LMM, β = 4.129%, $p < 0.001$).

Patient-related factors

Figure 4 displays the relationship between patient-related factors and segmentation performance. Male patients showed slightly better segmentation results than female patients, with a 1.0% higher DSC for males (LMM, β = 0.989%, $p < 0.001$). Older patients showed a decrease in DSC, but this effect was not significant (LMM, β = -0.011%, $p = 0.126$). We used the mean value of HU across the mandibular region to assess bone density and found that lower bone density reduced segmentation performance (Fig. 4c). The number of teeth in lower dentition positively influenced segmentation performance, with each additional tooth increasing DSC by 0.38% (LMM, β = 0.378%, $p < 0.001$). On the other hand, the presence of bone pathology (e.g. fractures, major cysts) reduced DSC by 0.71% (LMM, β = -0.708%, $p < 0.05$). Osteosynthesis material had the most significant negative effect, decreasing DSC by 7.90% (LMM, β = -7.90%, $p < 0.001$). Artifacts (e.g. metal, shadow) also negatively impacted segmentation, but showed no significant effect on DSC (LMM, β = -0.212%, $p = 0.3313$).

Anatomical Regions

Figures 5 and 6 visualize the case-wise segmentation using heatmaps. Most errors can be observed in the condyle, dentition, and part of the mandibular body. The segmentation performance of the AI model is significantly degraded in regions of impaired mandibular continuity (Case 21,65), bone pathology (Case 16,61), and osteosynthesis material (Case 17,86) (Supplementary Table 3). The segmentation results in Table 3 further demonstrate the differences in segmentation performance across Regions Of Interest (ROIs). The mandibular body performed the worst in terms of HD 95 and MASD. In terms of DSC, the condyle in CBCT had the lowest score of 78.07%. In addition, the dentition also had the lowest NSD value of 84.16%, indicating a lack of accurate boundary segmentation in this region. In summary, the mandibular body has the highest segmentation error in the distance-based metrics, whereas the condylar and dentition regions exhibit the lowest DSC and NSD, respectively.

Discussion

Although AI models have proven their performance, there are many open questions regarding the integration and limitations of current AI models in clinical routine as well as research. Recent qualitative research confirms that clinicians demand concrete insights into when and why AI fails in clinical settings, suggesting the need for comprehensive socio-technical evaluations[26]. Based on an experimental study with 20 current state-of-the-art AI models and the analysis of imaging features, patient characteristics, and anatomical regions on segmentation results, we were able to obtain new insights and provide recommendations for optimized social-technical setting, including clinical data acquisition and the requirements for future development of AI-based segmentation. To begin, our study required the creation of a benchmark dataset, as directly using public datasets or random sampling of private cases would not have been appropriate. Public datasets may overlap with the training data of the models under evaluation, and random sampling of private cases could not ensure a balance of imaging and patient features necessary for statistical analysis. Therefore, we built our benchmark dataset based on real-world scenarios where AI models are applied to end users, and determined the required size with a sample size calculation. Previous studies have shown that resampling could simulate multiple CBCT/CT scans from the same patient in a different image reconstruction settings[27]. Rotational movements of the patient's head could also be simulated using resampling methods[28]. Hence, we have created a quasi-experiment setting by resampling original CT/CBCT scans and manual screening of patient characteristics. This method provides enough data for the LMM to reveal the underlying factors influencing the performance of AI models.

Regulations on AI models

Among the 20 models selected for this study, the overall segmentation performance of the commercial models that had received MDR/FDA approval was higher than that of the research models (Fig. 2d). This suggests a positive impact of regulatory policies on the commercial model development and deployment process. However, the costs associated with certifying software as a medical device could be substantial. Regardless of the type of model, monitoring post-deployment performance is a critical step in improving safety as well as the effectiveness of AI models in clinical practice[29]. This is also a key feature of the overall product lifecycle approach used by the FDA[30]. As our study demonstrates, end-users should expect degradation in the performance of current static AI models as a result of changes in imaging protocols or changes in patient populations. One possible solution is dynamic fine-tuning of deployed models. However, the changes in performance as well as risk associated with this continuous learning may cause the product's metrics to differ from those at the time of initial certification, which would pose a significant regulatory challenge[31]. While regulators are actively developing guidance policies for dynamic tuning models, all approved AI tools have been static up to this date[32,33]. Therefore, the optimization of image acquisition protocols may be a viable alternative solution on static models. Furthermore, the identification of patient characteristics and anatomical regions that cause performance declines could lead to a strategy for intervening, both in the development of AI models as well as in their application.

Imaging factors and modality

The first questions arise in the optimal reconstruction protocol during the acquisition of medical imaging. Our investigation of one of the most versatile human bones, the mandible, suggests several key areas affecting the quality of AI-based bone segmentation. Elevated sharpness, decreased voxel size, and ensuring standardized patient positioning can all improve AI-based segmentation to a certain degree (Fig. 3). The results are in accordance with findings from traditional segmentation algorithms. Puggelli et al.[34] reconstructed CT scans of porcine tibiae with different kernels and evaluated the segmentation accuracy compared to laser scanning. The results demonstrated that sharp reconstruction kernel accuracy was higher than that of the soft kernel. The reason for that may be because the bone-soft tissue boundary is better defined in these images. Similarly, another study based on the segmentation results on CBCT scans of an AI model of 11 dry mandibles with different voxel sizes, revealed that larger voxels (0.45 mm) resulted in significant segmentation errors compared to smaller voxels (0.15 mm) (surface scans as reference)[35]. In contrast, Huang et al. concluded when applying one single AI model onto 183 CT scans of 11 patients with different voxel sizes, slice thickness and simulated doses, that there is no need for a strict image resolution[19]. Our comprehensive analysis with 20 models, however, underlined that lower sharpness (increased blurriness) as well as larger voxel size may have a negative impact on segmentation performance. This should be considered in the reconstruction protocols when incorporating AI models.

Another important factor is bone rotation during scanning (in our case the mandible). El Bachaoui et al. collected a total of 20 CBCT scans from 5 fresh cadavers at four different positions[36]. They concluded that the effect of sagittal rotation of the head on segmentation accuracy is clinically negligible (manual segmentation as reference). However, this study investigated a limited range of rotations in the sagittal plane only. In contrast, our study included a wide range of combined rotations in all three reference planes. Our results show that bone rotation in the axial and sagittal planes negatively affects the segmentation results (Fig. 3). This finding is probably due to the underlying distribution of the training data used by AI models. Attention should be paid to the standard positioning of the mandible, especially during CT scanning, as there is more freedom of movement for mandible on supine CT scans that lack chin fixation compared to CBCT. If a proper bone positioning cannot be achieved, post processing into a normalized bone position should be considered.

Regarding the imaging modality, most of the models trained with single modal data (CBCT or CT) were also able to segment scans of the other modality. Only one model, which trained solely on CT data, was unable to do so, as it successfully extracted the skull but was unable to separate the mandible from it. Such results indicate that CBCT and CT are interchangeable in this task, likely due to their similar fundamental imaging principles. Nevertheless, AI segmentation on CBCT demonstrated higher accuracy in descriptive statistics, but the AI model was even better at segmenting the CT data in LMM analysis which took multiple factors into account. The main reason for this may be that the original voxel size of CBCT (0.268 mm in average) is smaller than that of CT (0.442 mm in average), and smaller voxels size leads to better segmentation (Fig. 7). Another reason could be the anisotropy of CT voxels, i.e., slice thickness is generally not equal to in-plane voxel size. In previous studies, this negative effect was predominantly observed in the inter-slice direction, with the main areas affected including the cranial side of the condyle, the inferior border of the mandible, and the alveolar ridge, which is also observed in our study[17]. In contrast, LMM considers voxel size and slice thickness as independent factors, avoiding the interference of voxel morphology on modality. In conclusion, the use of high-resolution CT scans with isotropic voxels may further improve bone segmentation results of AI models.

Patient-related factors and Regions of Interests

Beside image-related factors, patient-related factors may also affect segmentation accuracy. Our results showed slightly better segmentation performance in males (LMM, β = 0.99%, $p < 0.001$) (Fig. 4). Yet this difference is marginal, it suggests that the AI models can be readily applied to both sexes. Interestingly, the presence of teeth improved segmentation results (for each additional tooth, LMM, β = 0.38%, $p < 0.001$). A possible explanation is that teeth act as extra anatomical landmarks for the AI models. Lacking teethless training data could also be a reason. Although restorations and implants are typically the source of artifacts, LMM analysis considered artifacts an individual factor, allowing our study to identify the impact of teeth on segmentation outcomes. However, bone pathology and osteosynthesis materials significantly reduced accuracy.

This result aligns to that from the study of Cui et al. of one single AI model, where evaluated on an external dataset of 407 CBCT scans, missing teeth (DSC, -0.8%), malocclusion (DSC, -0.9%), and metal artifacts (DSC, -2.0%) negatively affected segmentation results[37].

The accuracy of mandible segmentation varies in different anatomical regions (Table 3). The condyle exhibits lower accuracy, primarily due to its thin cortical bone and low density of cancellous bone, as well as the surrounding high-density cranial base structures. This results in lower contrast in the condylar region, especially in CBCT images[38]. This was confirmed by our LMM analysis across anatomical regions, where the segmentation performance of the condylar region in CT images is improved by 8.59% in DSC (LMM, β = 8.35%, $p < 0.001$) compared to CBCT images, while the improvement of the whole mandible segmentation is merely 4.1% (LMM, β = 4.13%, $p < 0.001$)(Fig. 7). The mandible body also exhibits a higher degree of error in segmentation, which may partially be attributed to the presence of artifacts from the crowns and brackets[39]. Another reason for the drop in the performance on the mandibular body is the discontinuity of the mandible, often accompanied by large osteosynthesis reconstruction plates (Fig. 5 Fig. 6). This could lead to a partial segmentation failure, which in turn severely affects the overall segmentation performance of the mandibular body.

Ideally, AI segmentation models should not be sensitive to reconstruction protocols, patient factors, and anatomical regions, which are highly variable in a socio-technical system. However, due to the limitations in architecture and training data, the current models have not yet reached this goal. Nevertheless, according to our findings, the segmentation performance of the model can be improved by optimizing the imaging protocol. Simulated calculation with results from LMMs suggested that with a recommended protocol (CT scan, sharpness of about 5000 HU/mm, voxel size of 0.5 mm, and neutral bone position), an increase of 9.02% in DSC for AI segmentation can be expected, comparing to the worst combination. In terms of patient characteristics, AI segmentation on a young male with complete dentition, without artifacts, pathology, or osteosynthesis, the DSC would increase by 16.59% compared to the worst combination of features. With these two aspects into account, the difference in DSC between the cases adapted most to fitting predicted requirements of AI models in general and those least adapted would be 25%. A real pair of examples can be found in our dataset (Case 21 and Case 78, Supplementary Table 2), where the mean DSC for AI segmentation of the original volume was 71.82% and 91.49%, respectively, with a difference of 19.67%. This 20% difference in DSC is substantial in terms of workload since cases with DSC above 90% require minor adjustment and those below 75% need intensive manual involvement (Figure 2a).

Recommendations and Requirements

To narrow this performance gap in clinical practice, collaboration between clinicians and AI developers must focus on mutual adjustments informed by real-world needs. Clinicians can optimize imaging protocols to align with current AI capabilities, while developers should prioritize the requirements that address recurring clinical challenges.

For clinicians, understanding the technical limits of AI models is critical. To improve bone segmentation outcomes, we recommend using CT scans with small, isotropic voxels (0.5 mm or smaller) and high-sharpness protocols when possible. In terms of modality, clinicians should be aware of the potential performance drop in susceptible regions like condyles in CBCT. Also, ensure target bones are positioned neutrally during scans, if not possible (e.g. trauma), adjust the images to a standard orientation before segmentation. In cases with edentulous mandible, large implants, or bone pathologies, clinicians should expect lower accuracy and prepare for manual corrections.

For AI developers, the next-gen models should be stable in performance even when faced with non-ideal clinical conditions. This includes robustness to patient features like bone pathology and osteosynthesis. Considering the sparsity of specific patient group, synthetics data can be a viable option. Segmentation performance in complex anatomical regions (e.g. condyles) should be prioritized, which could be achieved through regionally weighted loss functions or adversarial training for specific structures. In addition, models should explicitly flag uncertain or low-confidence segmentation regions by heatmaps or scores to guide clinician review, particularly in high-risk cases involving bone pathologies or surgical planning.

Limitation

Our study recruited the largest number of AI models to date and comprehensively analyzed the socio-technical factors including patient factors and imaging factors on segmentation performance. However, one limitation of the study is that we focused on bone segmentation only, which is only one but important fraction of the human anatomy. It would be interesting to see similar investigations into soft tissue segmentation (e.g. hearts, lungs and livers). This may involve analyzing the performance of AI models in various imaging modalities commonly used on soft tissue such as MRI or 3D ultrasound. The impact of factors such as tissue deformation, movement artifacts and inter-patient variability on segmentation results could be factors to be further assessed. In addition, our dataset did not include cases under the age of 18 years because they are not common cases for mandibular bone segmentation. This prevented us from fully capturing anatomical variability in all clinical situations, especially in patients who grow and develop during childhood and adolescence.

Future work

On our benchmark dataset, the current models still have a certain number of unsatisfying segmentation results, and clinicians need to refine them manually using various tools (Figure 2a). Integrating models with interactive tools (e.g., SAM[40] and MedSAM[41]) could streamline this “last mile” by allowing clinicians to correct errors via intuitive prompts. This study only briefly investigated the basic architecture used by the models, and due to confidentiality reasons, we were not able to examine in detail the configuration of the training parameters of each model. As a result, the impact of these technical specifications, in addition to the black-box characteristics of AI models, on segmentation accuracy is still not fully understood. Future research should explore these factors, potentially by collaborating to configure models and data in a controlled environment for further experiments.

Conclusion

This multi-center study shows that the performance of AI mandible segmentation is dynamically shaped by socio-technical factors, including imaging protocols, patient-specific factors and anatomical complexity. Two pillars are essential to the success of clinical translation of AI models: clinicians should adapt their workflows to the current limitations of AI, and developers must tackle the upcoming requirements that address persistent clinical challenges. For clinical teams, this means choosing high-resolution CT protocols when possible, ensuring standardized patient positioning and rechecking AI output in cases involving bone pathology or osteosynthesis. For AI developers, the requirements for the next-gen AI segmentation models are summarized from clinical failures. Models must remain robust to common clinical variabilities like rotation. Models should further improve the accuracy of error-prone anatomical regions (e.g., condyles) and provide intuitive uncertainty feedback to guide clinical reviews. These are not standalone checklists but interconnected obligations—only through this dual commitment can AI progress from a static algorithm and technocentric preoccupation to a trustworthy clinical ally in a socio-technical system.

Declaration

Author Contributions: Conceptualization, B.P. and K.X.; Methodology, K.X. and B.P.; software, K.X., M.C. and B.P.; validation, Y.L., A.F. and B.P.; formal analysis, All authors; investigation, K.X., B.P., L.G. and M.C.; resources, B.P., R.R., F.H., M.G., J.S., J.X., E.T., T.PA., M.B., N.S., R.T., G.D., C.W., N.V., P.V., Y.G., Z.X., J.B., A.R., T.F., A.L., R.C., S.V., R.I., S.R., D.M., C.S., T.X., S.B., S.N., O.K., S.Z., M.W., O.S., F.T., H.L., A.C. and T.PO.; data curation, K.X., L.G. and B.P.; writing—original draft preparation, K.X.; writing—review and editing, B.P., F.H., R.R., A.H., S.Z., A.L., and the rest of authors; visualization, K.X. and B.P.; supervision, B.P.; project administration, B.P.; funding acquisition, B.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The ethics application of the study was approved by the ethics committee at the RWTH Aachen University (approval number 23-272, 26th October 2023, Prof. Dr. Ralf Hausmann).

Informed Consent Statement: No informed consent was needed due to the use of anonymized retrospective patient data.

Data Availability: Due to the model anonymity nature of the study, only the evaluation result with code names of the model is made available in our repository. Benchmarking dataset and the model predictions are available on request from the corresponding author.

Code Availability: The code for dataset preparation and model evaluation were implemented in Python (Version 3.11.0). The source code and R code for statistical analysis is available on GitHub (https://github.com/OMFSdigital/AI_Mandible_Benchmarking).

Acknowledgments: We thank the anonymous patients whose CT and CBCT scans formed the basis of this study.

Conflicts of Interest: This research employs eight commercial AI models from companies. Some of the co-authors are employed by or have financial ties with these companies. Jan Schepers and Adriaan Lambrechts are employed by Materialise NV. Tobias Pankert and Stefan Raith are employed by Inzipio GmbH. Charlotte Weschke and Hans Lamecker are employed by 1000shapes. Ross Cotton is employed by Synopsys Northern Europe Ltd. Oldřich Kodym is employed by TESCAN 3DIM. Antoine Coppens is employed by Relu BV. Thomas Potrusil is employed by CADS GmbH and KLS Martin Group. Osku Sundquivst is employed by Planmeca Oy. It is important to note that the companies and institutions only provided model information and conducted inference on the benchmark dataset, without involvement in data analysis or evaluation results. In addition, model performance data have been anonymized for all authors (except for Kunpeng Xie and Behrus Puladi) using model designation codes. Despite these relationships, all necessary measures were taken during the study's design, data collection, and analysis to ensure the objectivity and integrity of the research findings. All other authors declare no conflicts of interest.

Tables

Table 1. AI models Summary

Name	Institute/Company	Location	Architecture
AC-Seg[9]	Inzipio GmbH	Aachen, Germany	3D-UNet[42]
SegCBCT	University Zurich	Zurich, Switzerland	3D-UNet
SKUBA	CADS	Perg, Austria	nnUNet[43]
MandibleSegNet	Charité University Medicine/ZIB	Berlin, Germany	nnUNet
3D-JMax[8]	University Hospital Basel	Basel, Switzerland	3D-UNet
Mandible	1000shapes	Berlin. Germany	nnUNet
JawFracNet[44]	Radboudumc	Nijmegen, The Netherlands	3D-UNet
JLU-Mandible	Jilin University	Changchun, China	U-Mamba[45]
Relu Creator[7]	Relu BV	Leuven, Belgium	3D-UNet
MandiSeg-Swin	IKIM Essen	Essen, Germany	SwinUNETR[46]
DentalSegmentator[47]	Arts et Métiers Institute of Technology	Paris, France	nnUNet
nnHaN-Net[48]	UMIT TIROL	Tirol, Austria	nnUNet
Planmeca Romexis Smart Lite	Planmeca	Helsinki, Finland	DynUNet[49]
FastJaw	TESCAN	Czech Republic	Cascaded U-nets
Simpleware CMF	Synopsys	California, USA	DNN
Materialise CMF segmentation model	Materialise NV	Leuven, Belgium	Confidential
KAAK	UMCG	Groningen, The Netherlands	nnUNet[50]
Edge Supervison Segmentation[50]	SJTU	Shanghai, China	3D-VNet[51]
AMASSS-CBCT[52]	University of Michigan (Public)	Michigan, USA	3D-UNETR
MedLSAM[53]	OpenMedLab (Public)	Shanghai, China	MedSAM[41], MedLAM

Table 1. Summary of the recruited AI models.

Table 2. Performance Summary

Model	DSC (%)			NSD (%)			HD 95 (mm)			MASD (mm)		
	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall
A	**90.87±9.19**	**94.01±4.86**	**92.44±7.51**	**93.77±9.80**	**94.37±5.55**	**94.07±7.97**	4.05±20.70	**2.26±5.05**	**3.16±15.09**	**0.60±3.02**	**0.29±0.46**	**0.44±2.16**
B	87.82±7.03	92.11±5.48	89.97±6.66	89.56±9.84	91.93±7.02	90.75±8.62	4.71±11.64	3.90±10.74	4.31±11.20	0.64±1.17	0.60±1.78	0.62±1.50
C	88.60±7.22	90.57±8.87	89.59±8.15	91.34±9.60	91.90±9.60	91.62±9.60	7.29±28.00	5.08±12.31	6.18±21.64	1.71±11.48	0.72±1.99	1.21±8.25
D	82.62±18.74	89.81±11.81	86.24±16.03	83.40±20.21	89.50±12.03	86.48±16.87	19.90±56.52	4.68±9.96	12.23±41.12	3.15±11.08	0.70±1.76	1.91±7.99
E	88.19±6.92	87.91±12.35	88.05±10.01	89.18±9.48	85.71±13.66	87.44±11.88	10.23±33.46	11.32±23.69	10.77±28.98	1.45±5.52	1.77±4.24	1.61±4.92
F	86.16±8.05	92.37±6.80	89.27±8.07	88.60±10.81	93.62±8.52	91.11±10.04	5.98±13.37	3.08±9.73	4.53±11.78	0.78±1.82	0.48±1.51	0.63±1.68
G	88.57±4.81	89.02±10.82	88.79±8.37	90.04±7.00	87.57±10.76	88.80±9.16	6.70±14.40	8.13±11.86	7.41±13.20	0.62±0.95	0.95±2.41	0.78±1.84
H	85.94±12.05	88.63±9.26	87.29±10.81	87.80±13.99	87.76±10.85	87.78±12.50	12.65±29.67	9.47±17.75	11.06±24.49	2.81±11.15	1.15±2.27	1.98±8.09
I	86.81±8.61	87.77±13.56	87.29±11.35	88.66±12.07	88.65±15.19	88.65±13.70	4.68±11.40	9.26±18.64	6.96±15.59	0.61±0.91	1.32±3.11	0.96±2.31
J	88.43±3.98	90.22±4.61	89.32±4.39	90.45±6.78	89.95±6.30	90.20±6.55	6.92±22.98	5.92±7.42	6.42±17.07	0.74±2.20	0.61±1.04	0.68±1.72
K	80.92±18.31	80.63±13.45	80.78±16.06	83.16±18.25	80.64±12.93	81.90±15.86	16.74±26.51	10.19±9.70	13.47±20.22	2.15±4.49	1.33±1.35	1.74±3.34
L	84.45±10.07	85.28±13.85	84.86±12.11	86.58±12.82	83.54±13.65	85.06±13.32	9.17±21.49	7.70±13.45	8.44±17.93	1.19±2.62	1.20±2.34	1.20±2.48
M	85.64±4.10	87.45±7.56	86.55±6.15	87.12±10.02	86.13±9.80	86.63±9.92	**3.87±9.59**	5.18±12.65	4.52±11.24	0.61±0.69	1.02±3.22	0.82±2.34
N	82.02±15.89	49.38±34.36	69.13±29.55	83.12±16.03	47.03±32.13	68.87±29.56	17.48±61.10	52.16±37.07	33.29±54.35	2.90±13.85	16.60±16.09	9.14±16.40
O	80.74±8.61	83.61±9.50	82.17±9.17	80.76±12.62	82.72±10.81	81.74±11.79	37.07±47.93	10.63±16.60	23.88±38.24	5.12±7.57	1.41±2.83	3.27±6.01
P	79.72±11.64	82.02±10.83	80.87±11.30	83.31±10.90	78.50±9.30	80.91±10.41	6.48±15.68	7.38±15.00	6.93±15.34	0.86±1.95	1.09±2.49	0.98±2.24
Q	78.11±16.30	46.39±27.64	62.59±27.58	79.74±15.64	47.31±24.90	63.87±26.28	15.75±26.39	32.67±24.16	24.05±26.69	2.09±4.19	6.43±6.69	4.22±5.96
R	81.26±8.63	80.97±10.36	81.12±9.53	84.55±8.06	79.54±8.05	82.05±8.43	5.83±16.56	4.53±4.46	5.18±12.14	0.75±2.69	0.64±0.66	0.69±1.96
S	50.71±30.98	**NA**	50.71±30.98	50.78±30.43	**NA**	50.78±30.43	105.38±130.33	**NA**	105.38±130.33	21.47±29.62	**NA**	21.47±29.62
T	47.23±25.31	46.52±20.26	46.87±22.87	36.49±21.60	34.73±13.85	35.60±18.11	64.49±23.82	34.85±5.88	49.53±22.76	13.58±9.00	7.99±4.13	10.76±7.52
Overall	81.38±17.57	81.97±20.02	81.66±18.80	82.57±19.64	80.74±20.96	81.69±20.31	17.97±46.97	11.63±19.88	14.89±36.57	3.13±10.34	2.31±5.70	2.73±8.42

Table 2. Segmentation performance (mean ± sd) of AI models on the whole mandible. Best performances were marked in blue. Model S failed to segment CBCT volumes. Models anonymized by descending order of number of segmentations with DSC > 90%.

Table 3. Anatomical Region Summary

ROI	DSC (%)			NSD (%)			HD 95 (mm)			MASD (mm)		
	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall	CT	CBCT	Overall
Condyle	82.26 ± 19.19	**78.07 ± 23.54**	**80.26 ± 21.48**	90.46 ± 17.37	**82.24 ± 22.10**	86.53 ± 20.19	2.24 ± 4.31	2.71 ± 2.92	2.47 ± 3.72	0.43 ± 0.89	0.62 ± 0.84	0.52 ± 0.87
Dentition	**80.15 ± 19.31**	83.13 ± 18.10	81.60 ± 18.79	85.41 ± 18.94	82.83 ± 18.58	**84.16 ± 18.81**	7.99 ± 26.86	5.42 ± 9.56	6.74 ± 20.41	1.51 ± 6.01	1.11 ± 2.81	1.32 ± 4.74
IAN Foramen	82.19 ± 14.45	85.58 ± 13.28	83.82 ± 14.00	95.94 ± 9.19	96.46 ± 8.77	96.19 ± 8.99	1.09 ± 1.02	0.93 ± 0.52	1.01 ± 0.82	0.12 ± 0.23	0.12 ± 0.15	0.12 ± 0.19
Inferior Border	84.64 ± 16.84	85.31 ± 18.81	84.97 ± 17.83	87.19 ± 17.44	84.08 ± 19.48	85.66 ± 18.53	10.11 ± 35.85	8.24 ± 15.90	9.19 ± 27.88	1.80 ± 7.74	1.54 ± 4.02	1.67 ± 6.19
Mandible Body	80.51 ± 18.36	82.37 ± 21.13	81.41 ± 19.78	**85.02 ± 20.49**	85.39 ± 21.93	85.20 ± 21.20	**21.67 ± 60.00**	**10.23 ± 20.96**	**16.10 ± 45.76**	**4.30 ± 14.22**	**2.41 ± 6.67**	**3.38 ± 11.24**
Whole Mandible	81.38 ± 17.57	81.97 ± 20.02	81.66 ± 18.80	82.57 ± 19.64	80.74 ± 20.96	81.69 ± 20.31	17.97 ± 46.97	11.63 ± 19.88	14.89 ± 36.57	3.13 ± 10.34	2.31 ± 5.70	2.73 ± 8.42

Table 3. Performance of AI models (mean ± sd) on 5 anatomical regions and the whole mandible. Worst performances were marked in red.

Figures

Figure 1 - Workflow

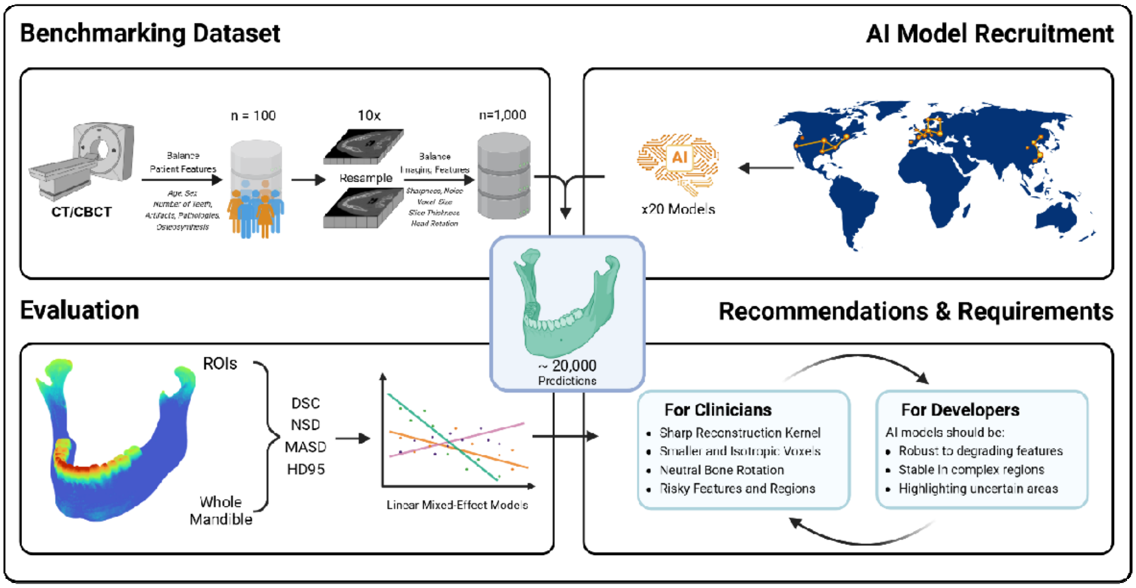


Figure 1. Workflow of the study. Created with BioRender.

Figure 2 - AI Models

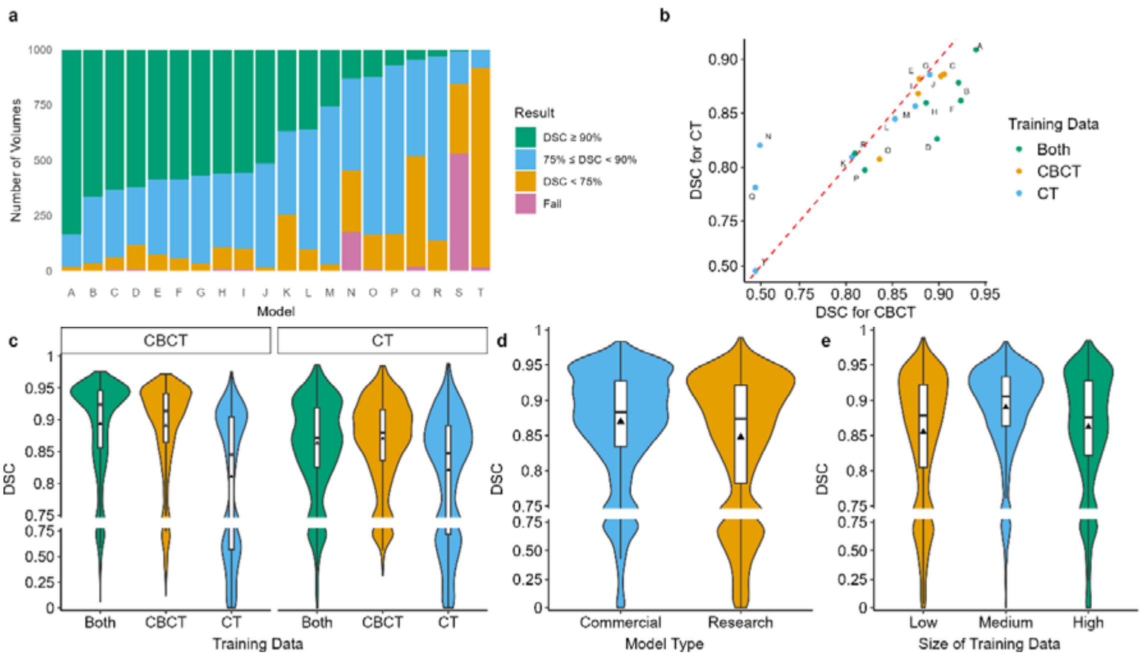


Figure 2. Model related factors and segmentation performance (**a**) Ranking of models based on segmentation quality. Decrease by number of good cases (DSC ≥ 0.9) (**b**) Distribution of model performance in CT and CBCT subsets based on mean DSC (**c**) Impact of training data on overall segmentation performance (**d**) Impact of model type (**e**) Impact of the size of training dataset. Low: 0-150 cases; Medium: 150-300 cases; High: 300+ cases.

Figure 3 - Image Quality

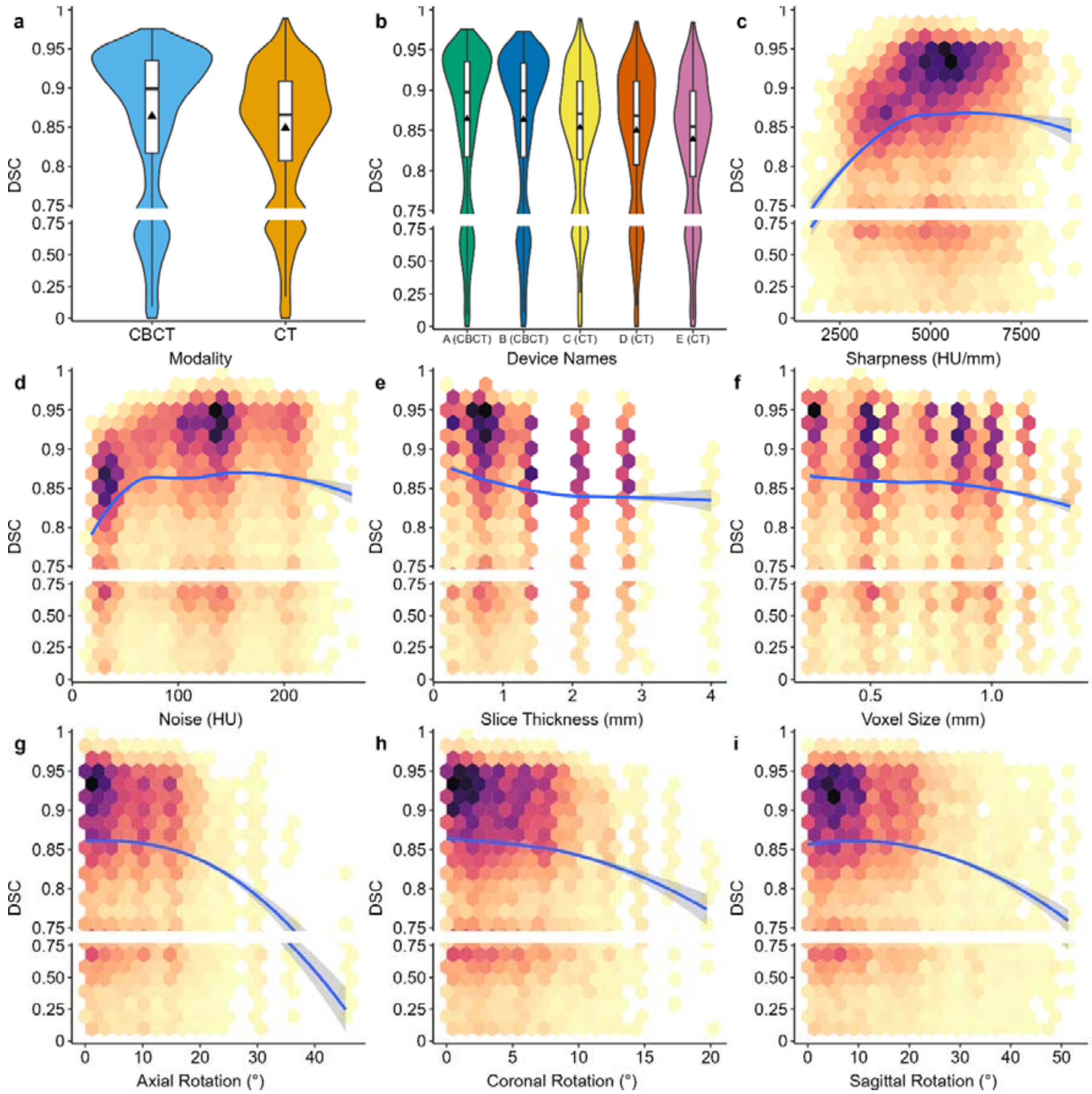


Figure 3. Image quality related factors and segmentation performance measured in DSC (**a**) distribution of segmentation performance in CBCT and CT scans (**b**) segmentation performance in five devices used in the study (**c**) relationship between image sharpness and segmentation performance (**d**) the effect of image noise image noise and segmentation performance (**e**) relationship between slice thickness and segmentation performance (**f**) the impact of voxel size on segmentation performance (**g**) ~ (**i**) the effect of bone rotation on segmentation performance. Colored hexagonal bins represent the distribution of data points. Darker colors indicate higher data density, while brighter colors indicate lower data density.

Figure 4 - Patient Characteristics

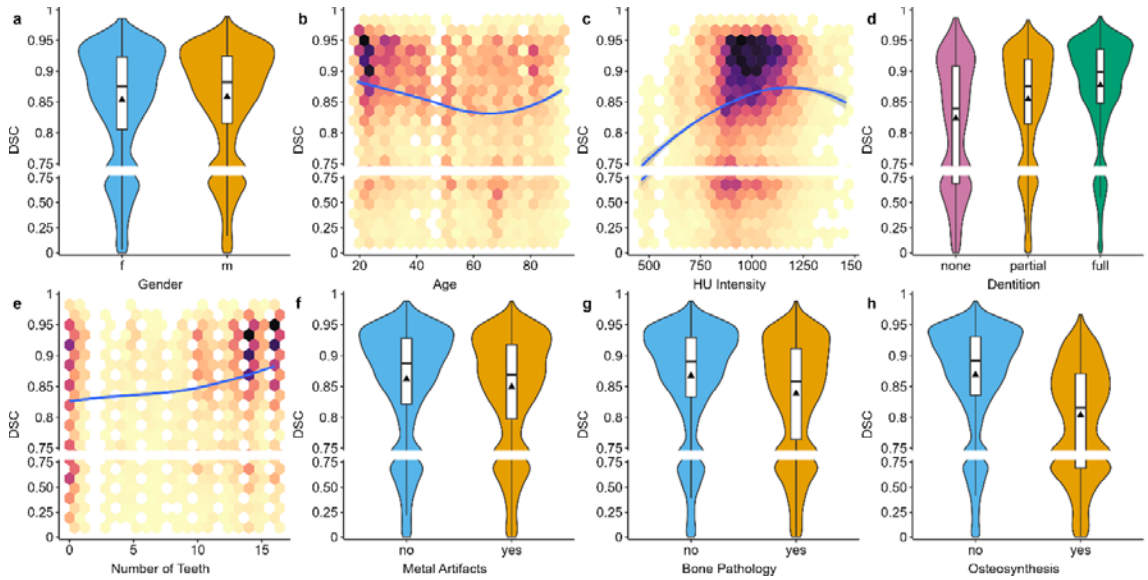


Figure 4. Patient related factors and segmentation performance (**a**)Comparison of between female and male patients (**b**) relationship between age and segmentation performance (**c**)The effect of Hounsfield Unit (HU) intensity on segmentation performance (**d**)(**e**) The impact of dentition status and teeth count on segmentation performance **(f)** Comparison of segmentation performance between cases with and without metal artifacts **(g)** Influence of bone pathology on segmentation performance **(h)** The effect of osteosynthesis on segmentation performance

Figure 5 - CBCT

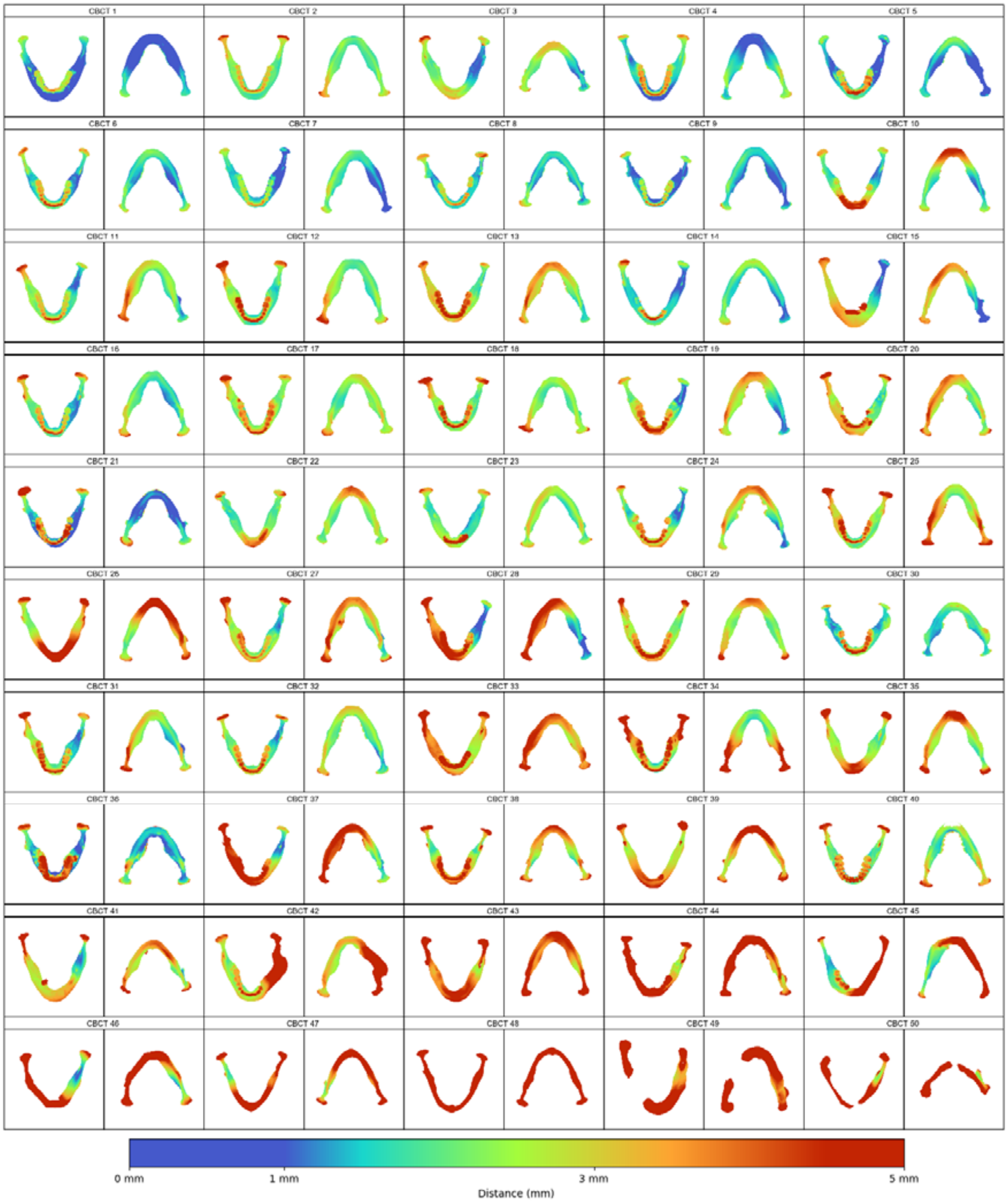


Figure 5. Heatmaps showing the average surface distance between AI segmentation results and the ground truths of CBCT scans. These segmentations were performed on the original scan and the 9 resample variants by 19 models (failed in model S), resulting in around 190 segmentations per case. Cases arranged in descending order of overall mean DSC.

Figure 6 - CT

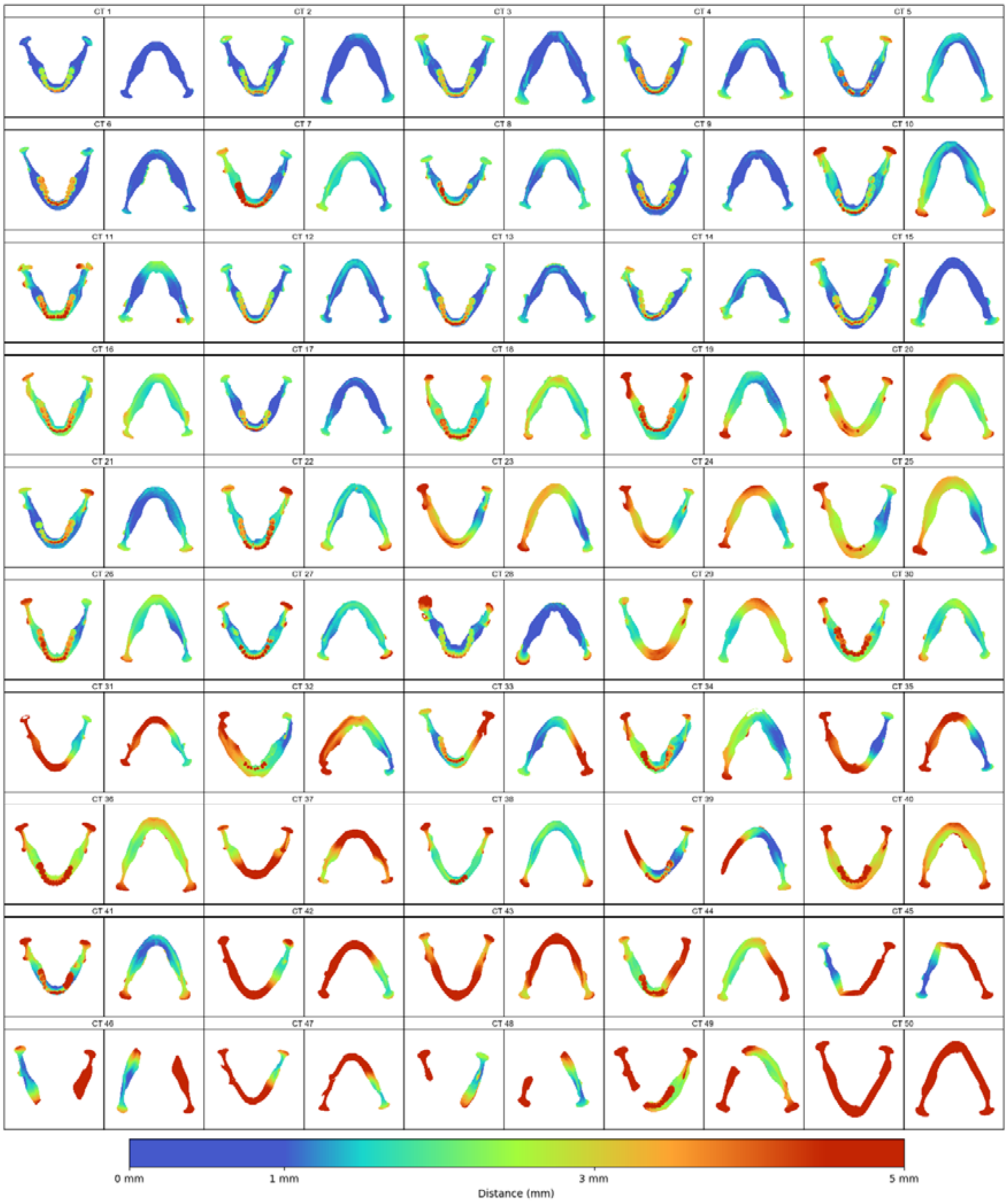


Figure 6. Heatmaps showing the average surface distance between AI segmentation results and the ground truths of CBCT scans. These segmentations were performed on the original scan and the 9 resample variants by 20 models, resulting in around 200 segmentations per case. Cases arranged in descending order of overall mean DSC.

Figure 7 - LMMs Summary in DSC

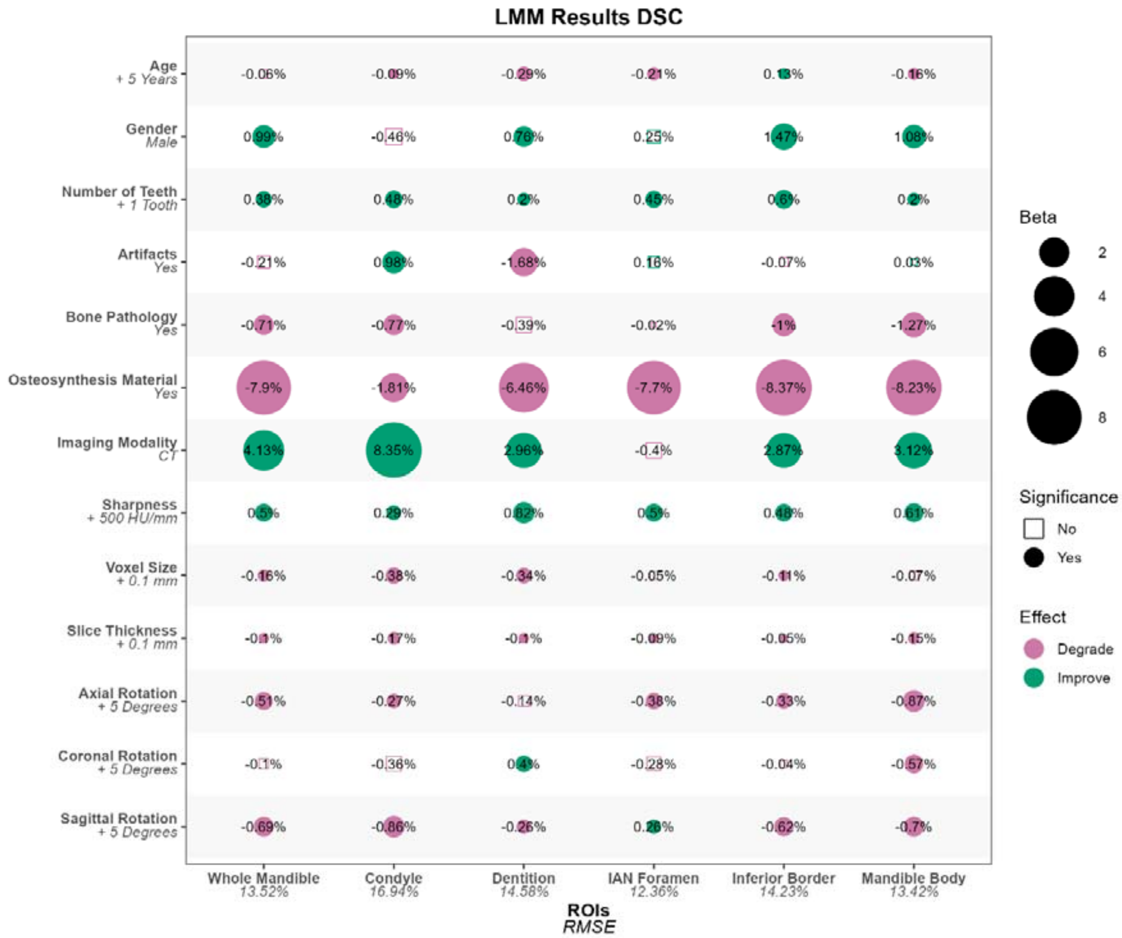


Figure 7. LMMs fitted on evaluation results in DSC% of five ROIs and the whole mandible. Factor considered significant when p<0.05.

Supplementary

Supplementary Table 1. Demographic and image features of the original scans

Patient features	CBCT (n=50)	CT (n=50)	Total (n=100)
Age			
Mean (SD)	46.240 (24.090)	50.700 (20.076)	48.470 (22.175)
Range	19 - 91	22 - 85	19 - 91
Gender			
f	25 (50.0%)	25 (50.0%)	50 (50.0%)
m	25 (50.0%)	25 (50.0%)	50 (50.0%)
Teeth			
Full	20 (40.0%)	14 (28.0%)	34 (34.0%)
None	10 (20.0%)	12 (24.0%)	22 (22.0%)
Partial	20 (40.0%)	24 (48.0%)	44 (44.0%)
Artifacts			
No	25 (50.0%)	27 (54.0%)	52 (52.0%)
Yes	25 (50.0%)	23 (46.0%)	48 (48.0%)
Bone Pathology			
No	32 (64.0%)	28 (56.0%)	60 (60.0%)
Yes	18 (36.0%)	22 (44.0%)	40 (40.0%)
Osteosynthesis			
No	40 (80.0%)	40 (80.0%)	80 (80.0%)
Yes	10 (20.0%)	10 (20.0%)	20 (20.0%)
Imaging features	**CBCT (N=50)**	**CT (N=50)**	**Total (N=100)**
Voxel Size			
Mean (SD)	0.268 (0.019)	0.442 (0.075)	0.355 (0.103)
Range	0.250 - 0.287	0.289 - 0.662	0.250 - 0.662
Slice Thickness			
Mean (SD)	0.268 (0.019)	0.706 (0.042)	0.487 (0.222)
Range	0.250 - 0.287	0.700 - 1.000	0.250 - 1.000
Device Name			
A	25 (50.0%)	-	25 (25.0%)
B	25 (50.0%)	-	25 (25.0%)
C	-	22 (44.0%)	22 (22.0%)
D	-	14 (28.0%)	14 (14.0%)
E	-	14 (28.0%)	14 (14.0%)

Supplementary Table 1. Demographic and image characteristics of the original scans

Supplementary Table 2. Comparison between best and worst Case

Factors	**Case 21**	**Case 78**	**Factor Unit**	**Beta(%)**	**Effect(%)**
Gender	F	M	-	0.99	0.99
Age*	65-70	20-25	5 years	0.06	0.528
Modality	CBCT	CT	-	0.38	6.08
Teeth Count	0	16	1 tooth	0.21	0.21
Artifacts	YES	NO	-	0.70	0.7
Bone Pathology	YES	NO	-	7.89	7.89
Osteosynthesis	YES	NO	-	4.13	4.13
Sharpness	4010.43	5326.17	500 HU/mm	0.50	1.32
Voxel Size	0.29	0.45	0.10 mm	0.16	-0.26
Slice Thickness	0.29	0.70	0.10 mm	0.09	-0.37
Axial Rotation	1.14	-4.47	5.00°	0.51	-0.34
Coronal Rotation	-2.31	-0.17	5.00°	0.10	0.04
Sagittal Rotation	-13.82	12.97	5.00°	0.69	0.12
DSC%_Original	71.82	91.49			
Real_diff (DSC %)					19.67
Model_diff (DSC %)					21.04

Supplementary Table 2. Sample cases showing the best combination of imaging and patient features verses the worst combination. A decline of 19.67% in DSC was observed. *To avoid identification, age ranges were used. The age difference between the two cases was 44 years.

Supplementary Table 3. Case-wise summary

Attached: Supplementary Table 3-CASE_RANKING.xlsx

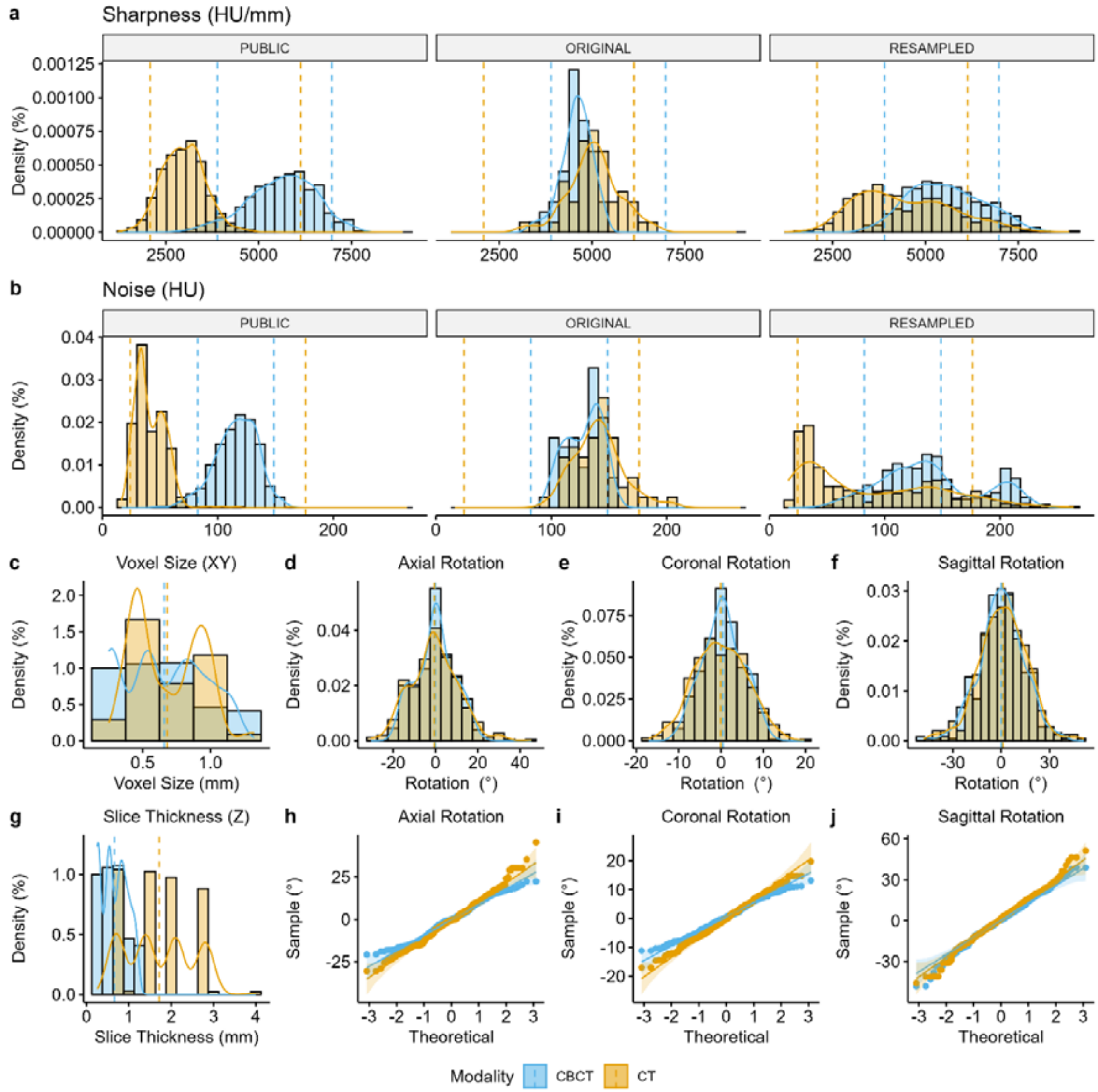
Supplementary Table 3. Average performance of the 20 AI segmentation models on the 100 original cases used in the study as well as their resampled versions for each case. The order of the cases is sorted by segmentation performance (DSC, HD95, MASD, NSD) from best to worst.

Supplementary Table 4. Resampling Factors

Attached: Supplementary Table 4-CASES_RESAMPLED_FINAL.xlsx

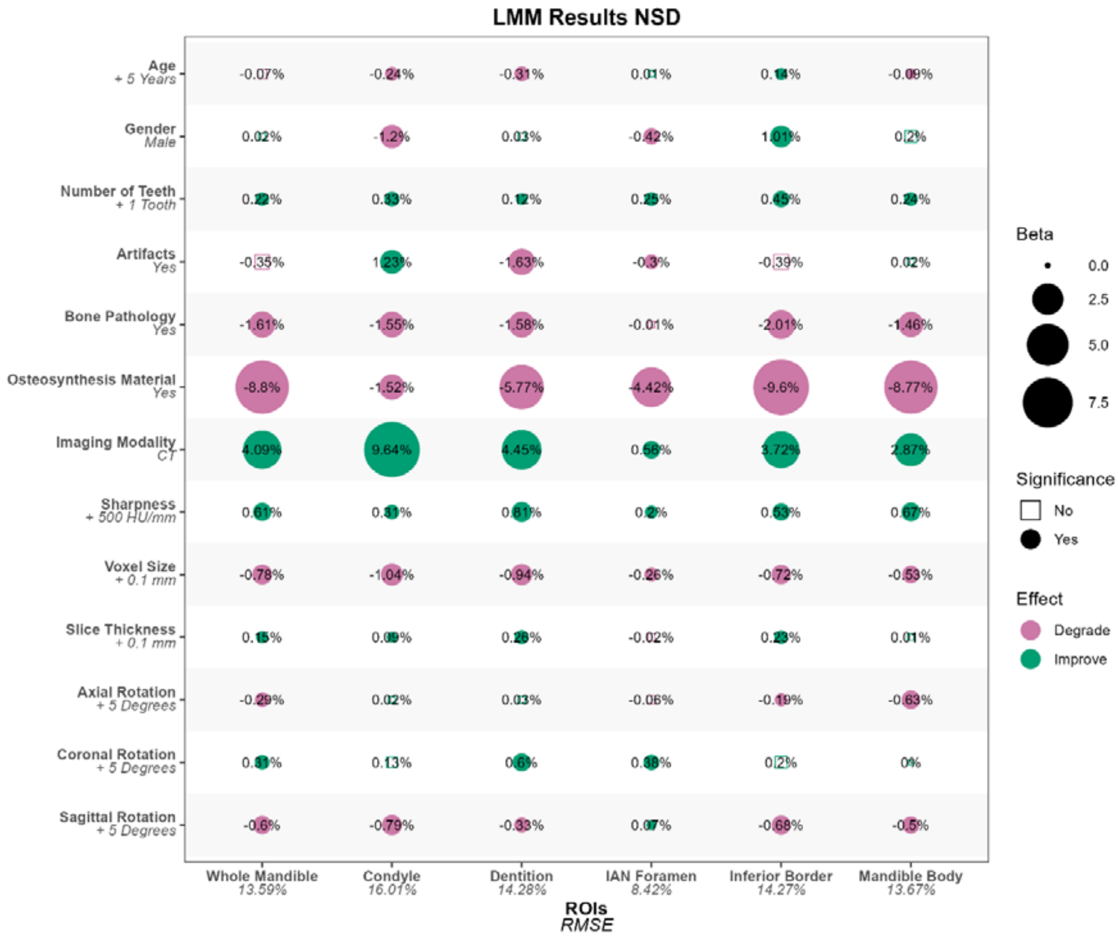
Supplementary Table 4. Resampling factors used for all 1000 volumes. The first 100 records are the original volumes. VOZ is the magnification of slice thickness and VXY is the magnification of in-plane voxel size. ROTX, ROTY, and ROTZ correspond to sagittal, coronal, and axial rotations, respectively. The columns SHARNESS and NOISE are measurements of sharpness and noise for that volume. See the online study protocol for more details in resampling.

Supplementary Figure 1. Distribution of imaging features in the final dataset



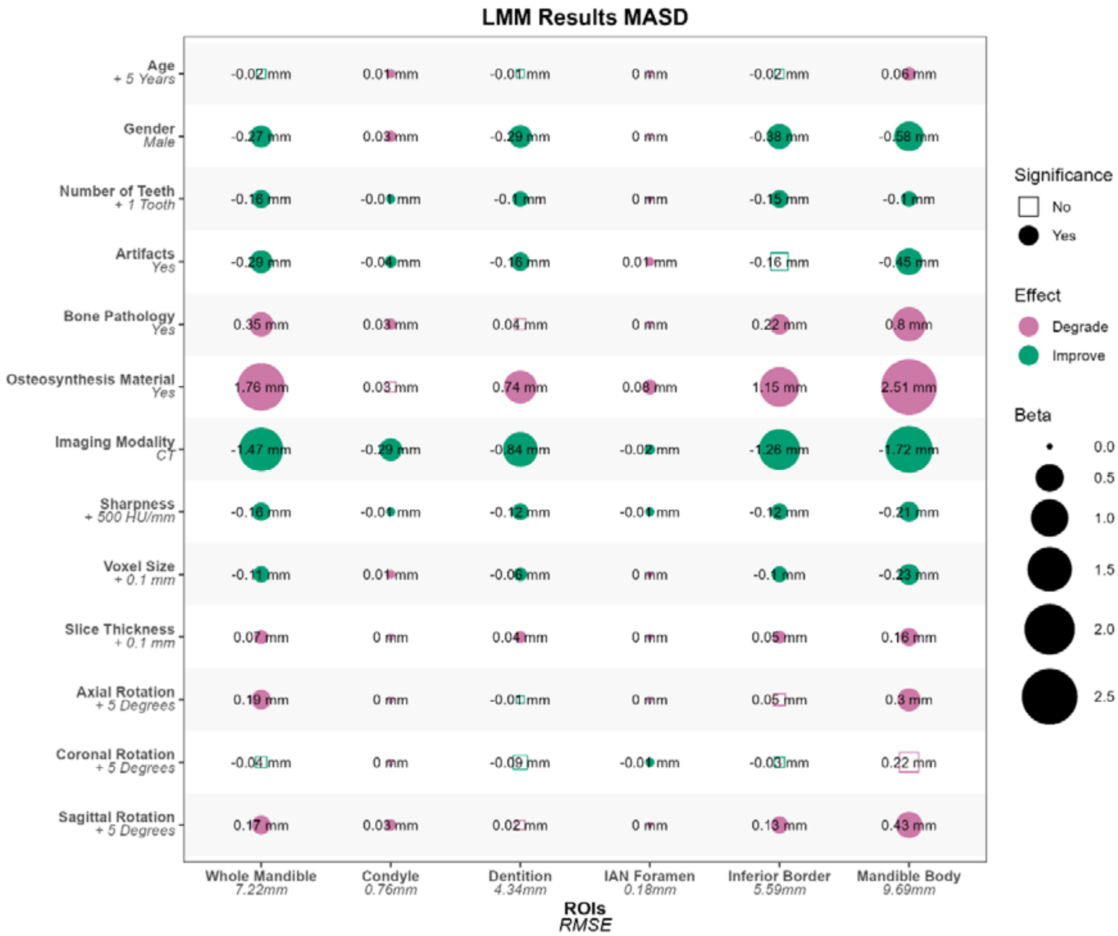
Supplementary Figure 1. Distribution of imaging features of the final dataset. a,b show the sharpness and noise distributions of the public dataset, the original scans, and the final dataset obtained from resampling, respectively. c and g present the overall voxel size and slice thickness of the final dataset. The final thickness of the CT is not more than 3 mm, and the CBCT voxels remain isotropic after scaling. d-f describe the distribution of the patient's mandible rotation angles in the final dataset. By adjusting the rotation parameters, the original minus mean value in the sagittal plane due to de-identified cropping have been compensated to approximately zero. h-i are Q-Q plots of the head rotation angle in the three planes, which show that the rotation angle variables are all close to a normal distribution.

Supplementary Figure 2. LMMs Summary in NSD



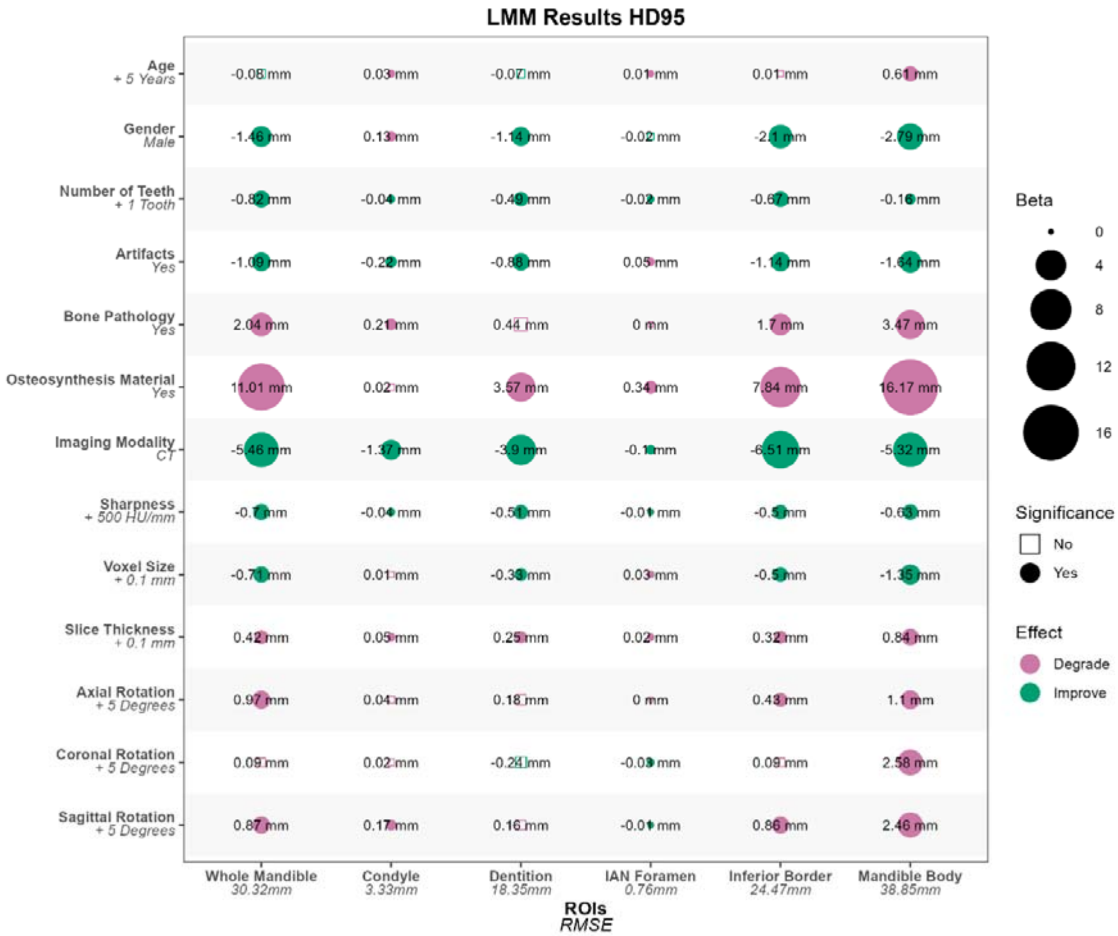
Supplementary Figure 2. LMMs fitted on evaluation results in NSD% of five ROIs and the whole mandible. Condyles are more affected by modality than in DSC metrics. Factor considered significant when p<0.05.

Supplementary Figure 3. LMMs Summary in MASD



Supplementary Figure 3. LMMs fitted on evaluation results in MASD (mm) of five ROIs and the whole mandible. Factor considered significant when $p<0.05$.

Supplementary Figure 4. LMMs Summary in HD95



Supplementary Figure 4. LMMs fitted on evaluation results in HD95 (mm) of five ROIs and the whole mandible. Factor considered significant when p<0.05.

References

1. van Baar, G. J., Forouzanfar, T., Liberton, N. P., Winters, H. A. & Leusink, F. K. Accuracy of computer-assisted surgery in mandibular reconstruction: A systematic review. *Oral Oncology* **84,** 52–60; 10.1016/j.oraloncology.2018.07.004 (2018).
2. Apostolakis, D., Michelinakis, G., Kamposiora, P. & Papavasiliou, G. The current state of computer assisted orthognathic surgery: A narrative review. *Journal of dentistry* **119,** 104052; 10.1016/j.jdent.2022.104052 (2022).
3. Fu, Y. *et al.* A review of deep learning based methods for medical image multi-organ segmentation. *1120-1797* **85,** 107–122; 10.1016/j.ejmp.2021.05.003 (2021).
4. van Eijnatten, M. *et al.* CT image segmentation methods for bone used in medical additive manufacturing. *Medical Engineering & Physics* **51,** 6–16; 10.1016/j.medengphy.2017.10.008 (2018).
5. Qiu, B. *et al.* Automatic Segmentation of Mandible from Conventional Methods to Deep Learning-A Review. *Journal of personalized medicine* **11**; 10.3390/jpm11070629 (2021).
6. Liu, P., Sun, Y., Zhao, X. & Yan, Y. Deep learning algorithm performance in contouring head and neck organs at risk: a systematic review and single-arm meta-analysis. *Biomedical engineering online* **22,** 104; 10.1186/s12938-023-01159-y (2023).
7. Verhelst, P.-J. *et al.* Layered deep learning for automatic mandibular segmentation in cone-beam computed tomography. *Journal of dentistry* **114,** 103786; 10.1016/j.jdent.2021.103786 (2021).
8. Ile□an, R. R., Beyer, M., Kunz, C. & Thieringer, F. M. Comparison of Artificial Intelligence-Based Applications for Mandible Segmentation: From Established Platforms to In-House-Developed Software. *Bioengineering (Basel, Switzerland)* **10**; 10.3390/bioengineering10050604 (2023).
9. Pankert, T. *et al.* Mandible segmentation from CT data for virtual surgical planning using an augmented two-stepped convolutional neural network. *International journal of computer assisted radiology and surgery* **18,** 1479–1488; 10.1007/s11548-022-02830-w (2023).
10. Zachow, S. Computational Planning in Facial Surgery. *Facial plastic surgery : FPS* **31,** 446–462; 10.1055/s-0035-1564717 (2015).
11. Antonelli, M. *et al.* The Medical Segmentation Decathlon. *Nat Commun* **13,** 4128; 10.1038/s41467-022-30695-9 (2022).
12. Rajamani, S. T. *et al.* Toward Detecting and Addressing Corner Cases in Deep Learning Based Medical Image Segmentation. *IEEE Access* **11,** 95334–95345; 10.1109/ACCESS.2023.3311134 (2023).
13. Maier-Hein, L. *et al.* Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun* **9,** 5217; 10.1038/s41467-018-07619-7 (2018).
14. Reddy, S. Explainability and artificial intelligence in medicine. *The Lancet. Digital health* **4,** e214-e215; 10.1016/S2589-7500（22）00029-2 (2022).
15. van de Sande, D. *et al.* To warrant clinical adoption AI models require a multi-faceted implementation evaluation. *NPJ digital medicine* **7,** 58; 10.1038/s41746-024-01064-1 (2024).
16. Whetton, S. & Georgiou, A. Conceptual challenges for advancing the socio-technical underpinnings of health informatics. *The open medical informatics journal* **4,** 221–224; 10.2174/1874431101004010221 (2010).
17. Gruber, L. J. *et al.* Accuracy and Precision of Mandible Segmentation and Its Clinical Implications: Virtual Reality, Desktop Screen and Artificial Intelligence. *Expert Systems with Applications* **239,** 122275; 10.1016/j.eswa.2023.122275 (2024).
18. Minnema, J. *et al.* Segmentation of dental cone-beam CT scans affected by metal artifacts using a mixed-scale dense convolutional neural network. *Medical physics* **46,** 5027–5035; 10.1002/mp.13793 (2019).

19. Huang, K. *et al.* Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. *Journal of applied clinical medical physics* **22,** 168–174; 10.1002/acm2.13207 (2021).

20. Wee, L. & Dekker, A. Data from HEAD-NECK-RADIOMICS-HN1, 2019.

21. Grossberg, A. *et al*. HNSCC, 2020.

22. Cipriano, M. *et al.* Deep Segmentation of the Mandibular Canal: A New 3D Annotated Dataset of CBCT Volumes. ToothFairy CBCT. *IEEE Access* **10,** 11500–11510; 10.1109/ACCESS.2022.3144840 (2022).

23. Nikolov, S. *et al*. Clinically Applicable Segmentation of Head and Neck Anatomy for Radiotherapy: Deep Learning Algorithm Development and Validation Study. *Journal of medical Internet research* **23,** e26151; 10.2196/26151 (2021).

24. Google DeepMind. *Surface distance metrics* (Github, 2022).

25. Maier-Hein, L. *et al*. Metrics reloaded: Recommendations for image analysis validation. *Nat Methods* **21,** 195–212; 10.1038/s41592-023-02151-z (2024).

26. Scott, I. A., van der Vegt, A., Lane, P., McPhail, S. & Magrabi, F. Achieving large-scale clinician adoption of AI-enabled decision support. *BMJ health & care informatics* **31**; 10.1136/bmjhci-2023-100971 (2024).

27. Naziroglu, R. E., van Ravesteijn, V. F., van Vliet, L. J., Streekstra, G. J. & Vos, F. M. Simulation of scanner- and patient-specific low-dose CT imaging from existing CT images. *Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics (AIFB)* **36,** 12–23; 10.1016/j.ejmp.2017.02.009 (2017).

28. Gardner, M., Bouchta, Y. B., Sykes, J. & Keall, P. J. A kinematics-based method for creating deformed patient-derived head and neck CT scans. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference* **2023,** 1–4; 10.1109/EMBC40787.2023.10340930 (2023).

29. Pesapane, F. *et al*. The translation of in-house imaging AI research into a medical device ensuring ethical and regulatory integrity. *European journal of radiology* **182,** 111852; 10.1016/j.ejrad.2024.111852 (2024).

30. FDA 2024. US FDA Artificial Intelligence and Machine Learning Discussion Paper.

31. Pianykh, O. S. *et al.* Continuous Learning AI in Radiology: Implementation Principles and Early Applications. *Radiology* **297,** 6–14; 10.1148/radiol.2020200038 (2020).

32. Article 96: Guidelines from the Commission on the Implementation of this Regulation | EU Artificial Intelligence Act. Available at https://artificialintelligenceact.eu/article/96/ (2024).

33. Marketing Submission Recommendations for a Predetermined Change Control Plan for Artificial Intelligence-Enabled Device Software Functions. *FDA* (Tue, 2024).

34. Puggelli, L., Uccheddu, F., Volpe, Y., Furferi, R. & Di Feo, D. Accuracy Assessment of CT-Based 3D Bone Surface Reconstruction. In *Advances on mechanics, design engineering and manufacturing,* edited by F. Cavas-Martínez, *et al.* (Springer Berlin Heidelberg, New York NY, 2019), pp. 487–496.

35. Alrashed, S., Dutra, V., Chu, T.-M. G., Yang, C.-C. & Lin, W.-S. Influence of exposure protocol, voxel size, and artifact removal algorithm on the trueness of segmentation utilizing an artificial-intelligence-based system. *Journal of prosthodontics : official journal of the American College of Prosthodontists* **33,** 574–583; 10.1111/jopr.13827 (2024).

36. El Bachaoui, S. *et al*. The impact of CBCT-head tilting on 3D condylar segmentation reproducibility. *Dento maxillo facial radiology* **52,** 20230072; 10.1259/dmfr.20230072 (2023).

37. Cui, Z. *et al*. A fully automatic AI system for tooth and alveolar bone segmentation from cone-beam CT images. *Nat Commun* **13,** 2096; 10.1038/s41467-022-29637-2 (2022).

38. Cuadros Linares, O., Bianchi, J., Raveli, D., Batista Neto, J. & Hamann, B. Mandible and skull segmentation in cone beam computed tomography using super-voxels and graph clustering. *Vis Comput* **35,** 1461–1474; 10.1007/s00371-018-1511-0 (2019).

39. Hirschinger, V., Hanke, S., Hirschfelder, U. & Hofmann, E. Artifacts in orthodontic bracket systems in cone-beam computed tomography and multislice computed tomography. *Journal of orofacial orthopedics = Fortschritte der Kieferorthopadie : Organ/official journal Deutsche Gesellschaft fur Kieferorthopadie* **76,** 152-60, 162-3; 10.1007/s00056-014-0278-9 (2015).

40. Ravi, N. *et al.* SAM 2: Segment Anything in Images and Videos, 2024.

41. Ma, J. *et al.* Segment anything in medical images. *Nat Commun* **15,** 654; 10.1038/s41467-024-44824-z (2024).

42. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. In *Medical Image Computing and Computer-Assisted Intervention {u2013} MICCAI 2016. 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II,* edited by S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal & W. Wells (Springer International Publishing; Imprint: Springer, Cham, 2016), pp. 424–432.

43. Isensee, F., Jaeger, P. F., Kohl, S. A. A., Petersen, J. & Maier-Hein, K. H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18,** 203–211; 10.1038/s41592-020-01008-z (2021).

44. van Nistelrooij, N. *et al.* Detecting Mandible Fractures in CBCT Scans Using a 3-Stage Neural Network. *Journal of dental research* **103,** 1384–1391; 10.1177/00220345241256618 (2024).

45. Ma, J., Li, F. & Wang, B. U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation, 2024/1/9.

46. Hatamizadeh, A. *et al.* Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images, 2022/1/4.

47. Dot, G. *et al.* DentalSegmentator: Robust open source deep learning-based CT and CBCT image segmentation. *Journal of dentistry* **147,** 105130; 10.1016/j.jdent.2024.105130 (2024).

48. Tappeiner, E., Welk, M. & Schubert, R. Tackling the class imbalance problem of deep learning-based head and neck organ segmentation. *International journal of computer assisted radiology and surgery* **17,** 2103–2111; 10.1007/s11548-022-02649-5 (2022).

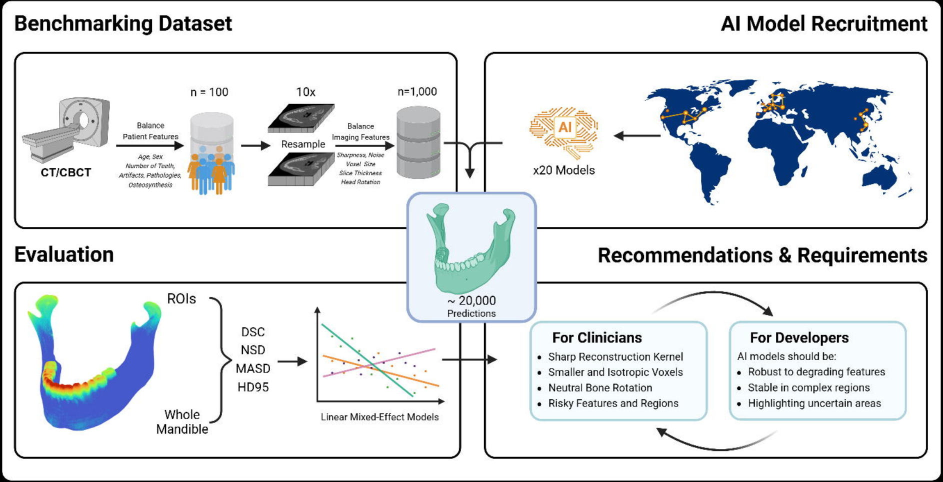
49. Ranzini, M. B. M., Fidon, L., Ourselin, S., Modat, M. & Vercauteren, T. MONAIfbs: MONAI-based fetal brain MRI deep learning segmentation, 2021/3/21.

50. Xu, J. *et al.* A 3D segmentation network of mandible from CT scan with combination of multiple convolutional modules and edge supervision in mandibular reconstruction. *Computers in biology and medicine* **138,** 104925; 10.1016/j.compbiomed.2021.104925 (2021).

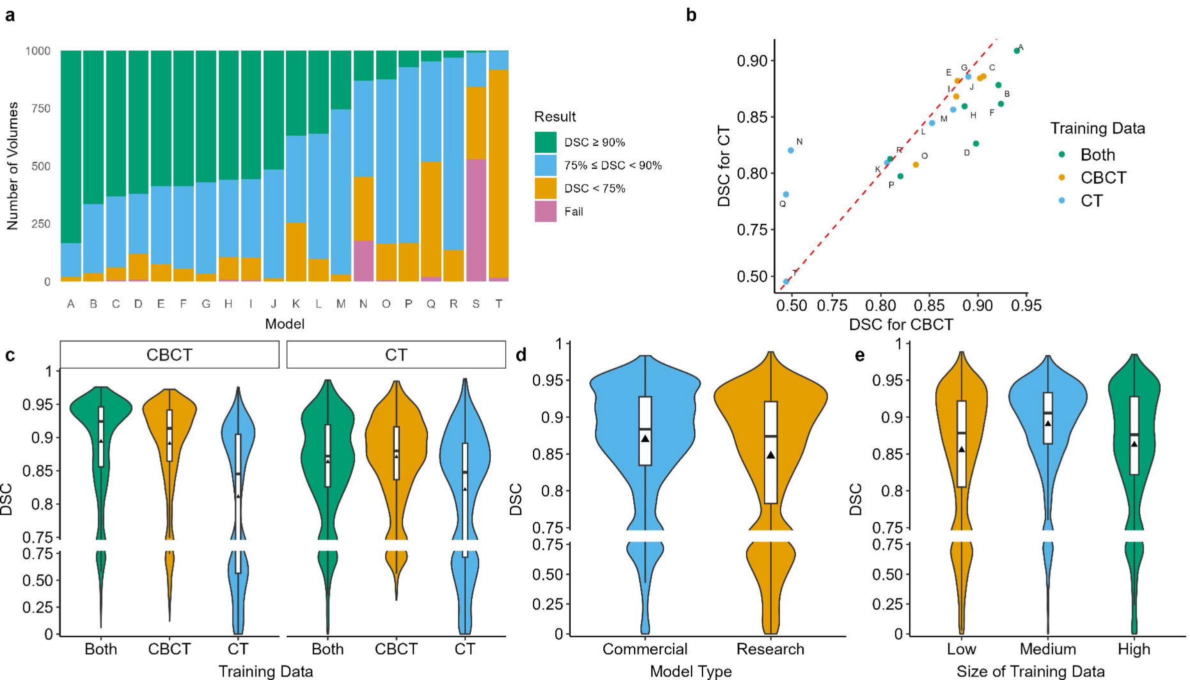
51. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)* (IEEE2016), pp. 565–571.

52. Gillot, M. *et al.* Automatic multi-anatomical skull structure segmentation of cone-beam computed tomography scans using 3D UNETR. *PloS one* **17,** e0275033; 10.1371/journal.pone.0275033 (2022).

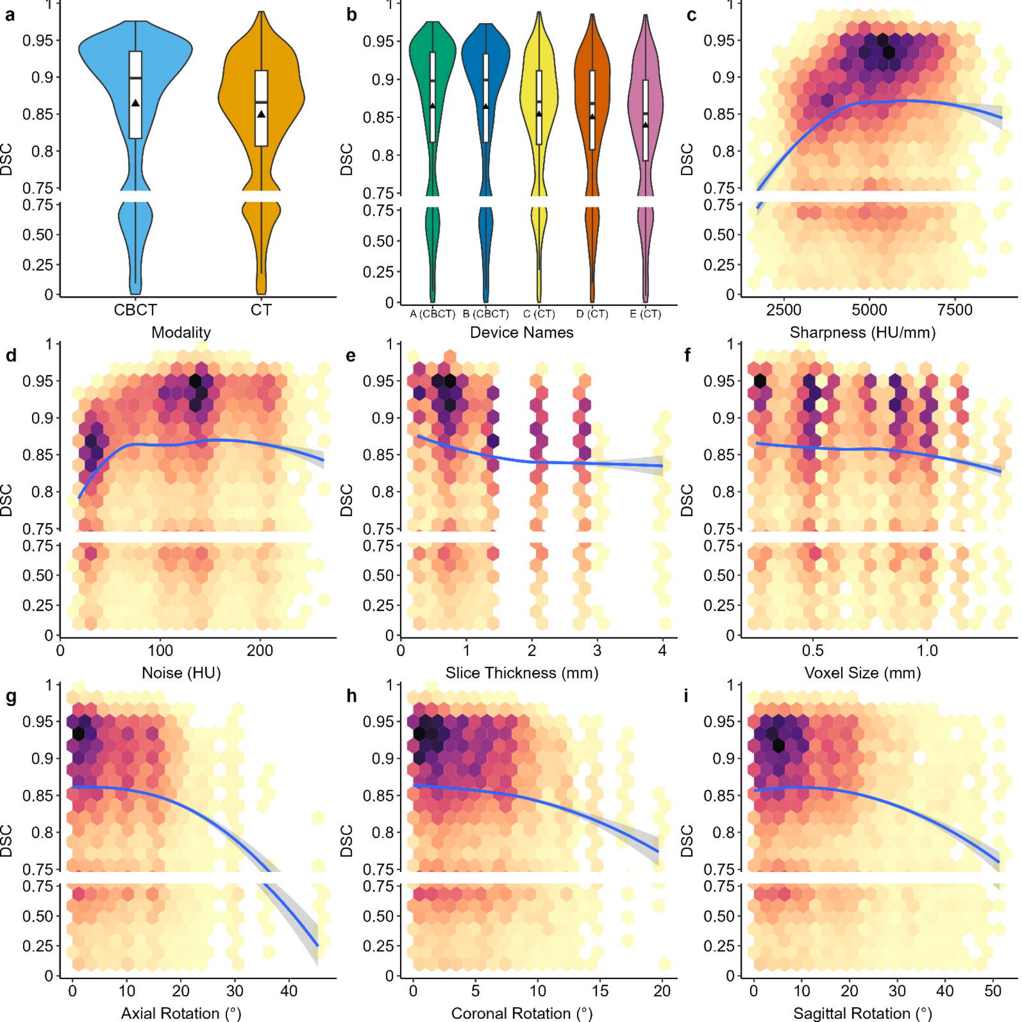
53. Lei, W., Xu, W., Li, K., Zhang, X. & Zhang, S. MedLSAM: Localize and segment anything model for 3D CT images. *Medical image analysis* **99,** 103370; 10.1016/j.media.2024.103370 (2025).



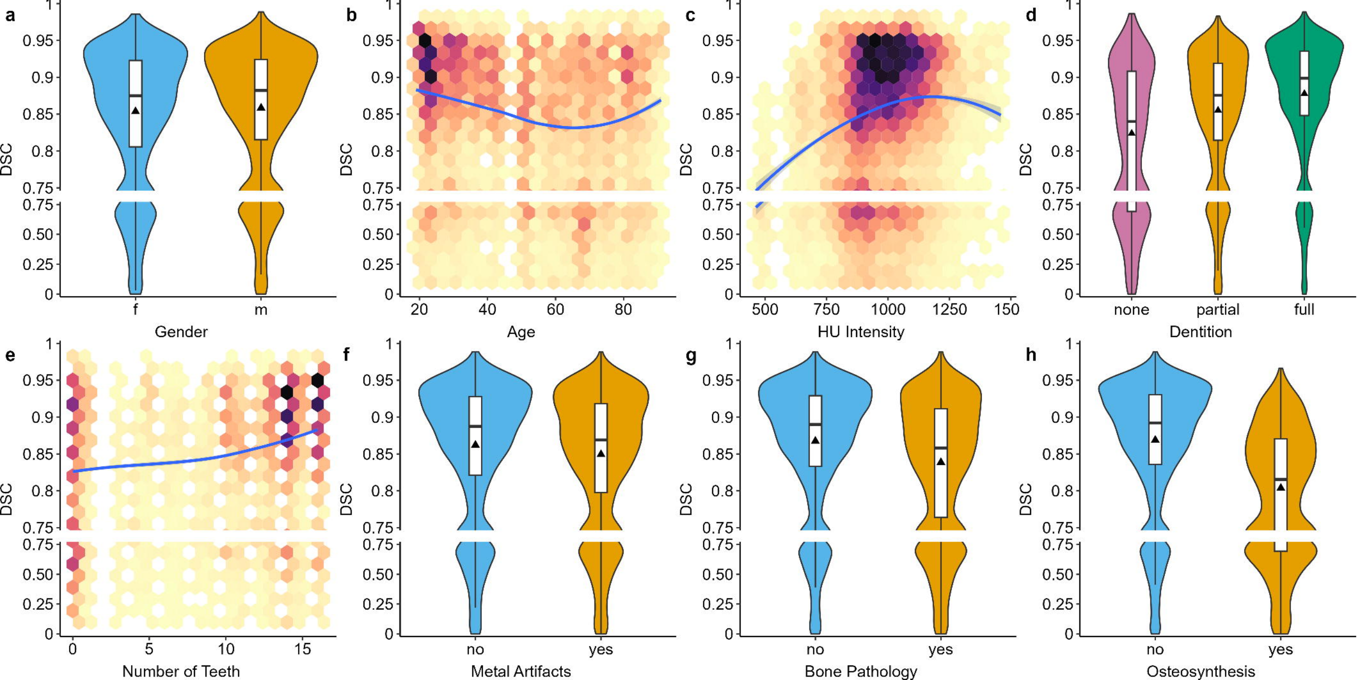
Benchmarking Dataset
AI Model Recruitment
CT/CBCT
Balance
Patient Features
Age, Sex
Number of Teeth,
Artifacts, Pathologies,
Osteosynthesis
n = 100
10x
Resample
Balance
Imaging Features
Sharpness, Noise
Voxel Size
Slice Thickness
Head Rotation
n=1,000
AI
x20 Models
~ 20,000
Predictions
Evaluation
Recommendations & Requirements
ROIs
Whole
Mandible
DSC
NSD
MASD
HD95
Linear Mixed-Effect Models
For Clinicians
Sharp Reconstruction Kernel
Smaller and Isotropic Voxels
Neutral Bone Rotation
Risky Features and Regions
For Developers
AI models should be:
Robust to degrading features
Stable in complex regions
Highlighting uncertain areas



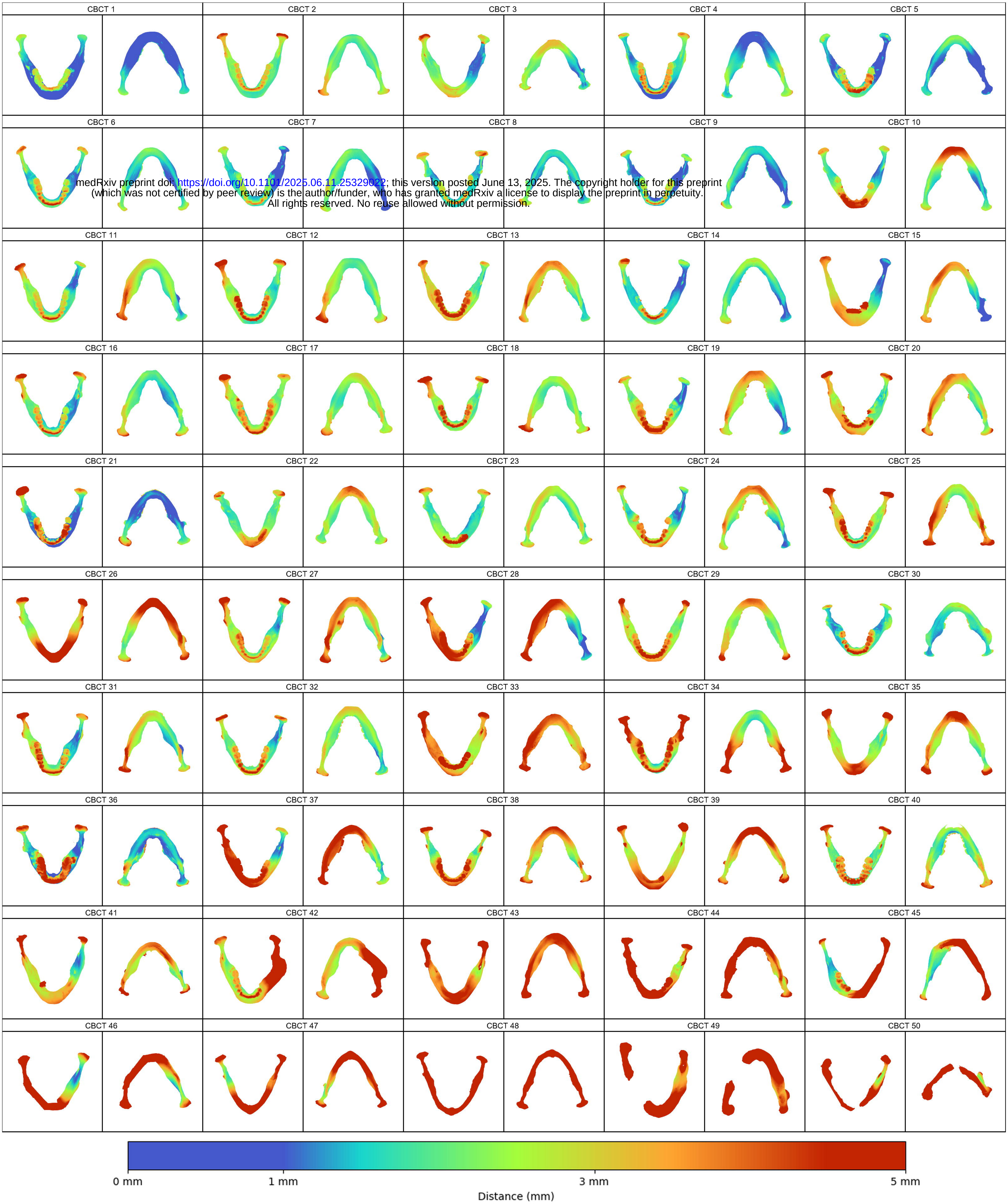
a
Number of Volumes
Model
Result
DSC ≥ 90%
75% ≤ DSC < 90%
DSC < 75%
Fail
b
DSC for CT
DSC for CBCT
Training Data
Both
CBCT
CT
c
CBCT
CT
DSC
Training Data
d
DSC
Commercial
Research
Model Type
e
DSC
Low
Medium
High
Size of Training Data

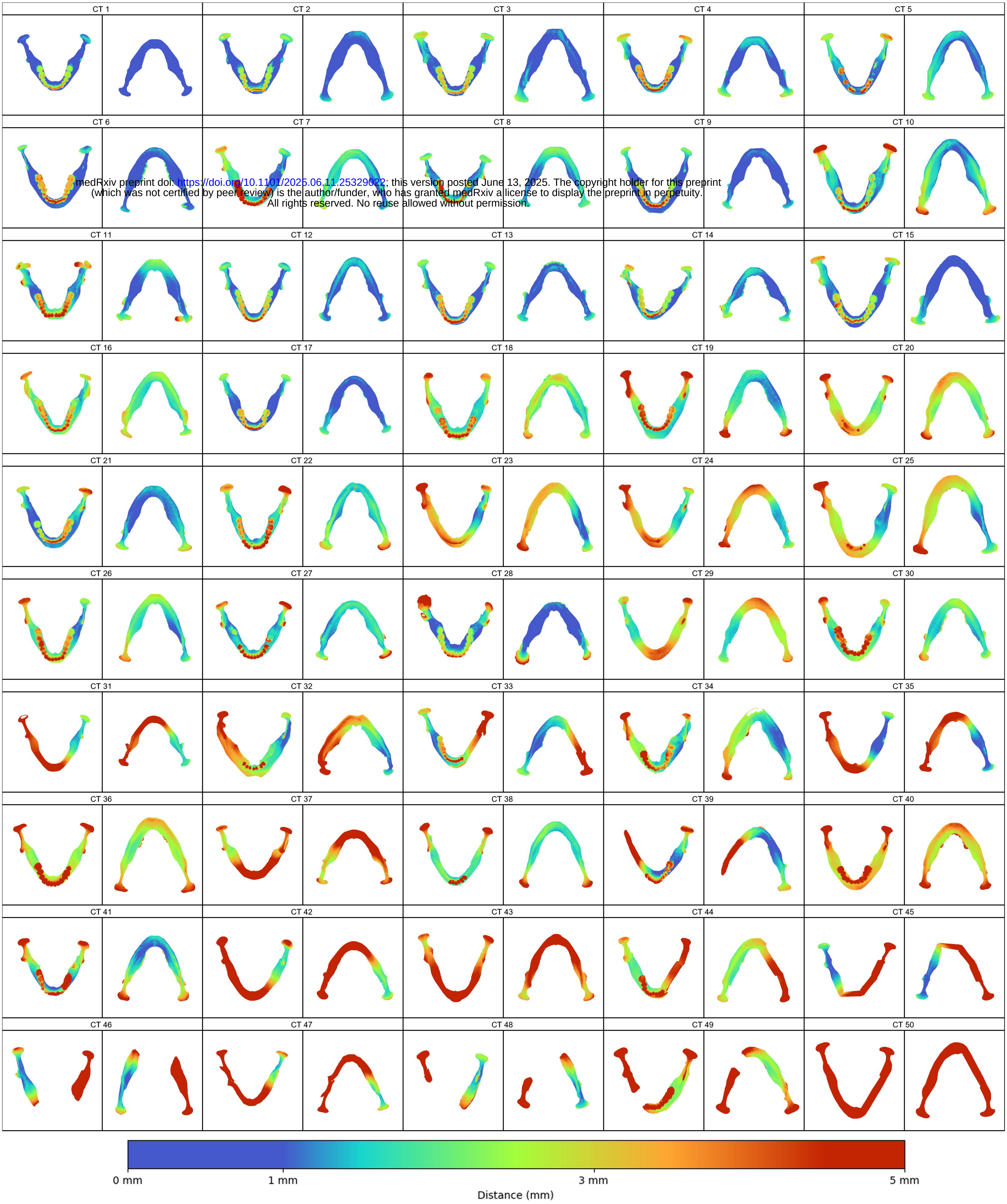


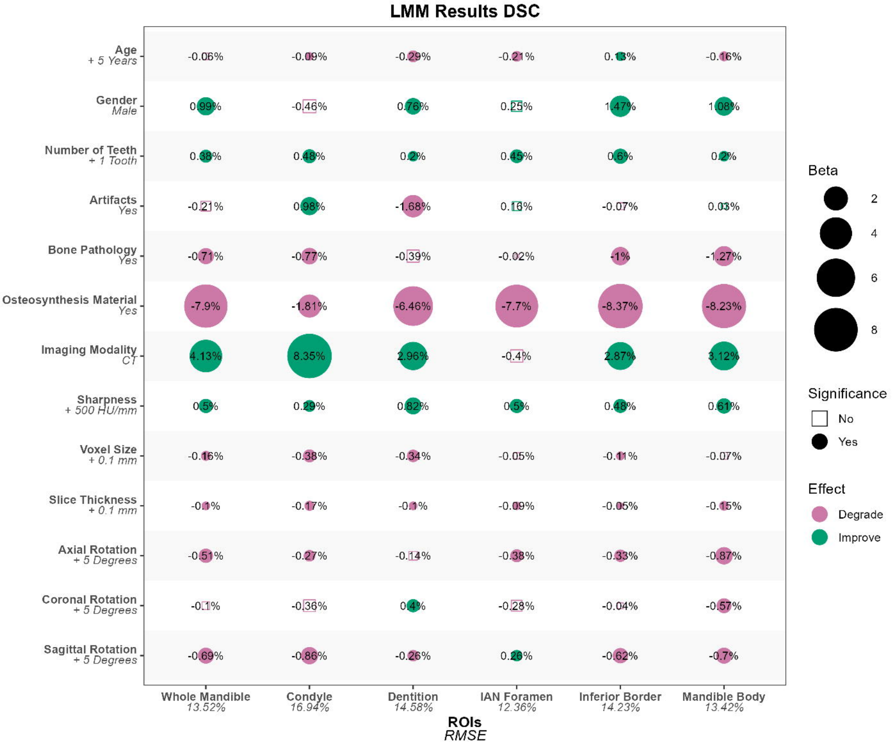
a
b
c
d
e
f
g
h
i
DSC
Modality
CBCT
CT
Device Names
A (CBCT)
B (CBCT)
C (CT)
D (CT)
E (CT)
Sharpness (HU/mm)
Noise (HU)
Slice Thickness (mm)
Voxel Size (mm)
Axial Rotation (°)
Coronal Rotation (°)
Sagittal Rotation (°)

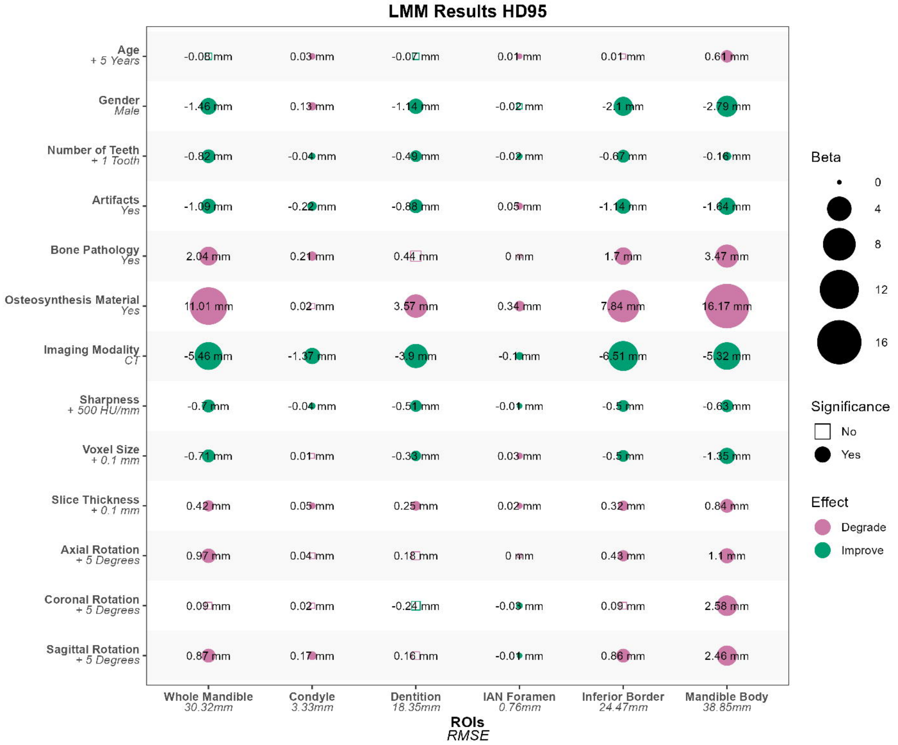


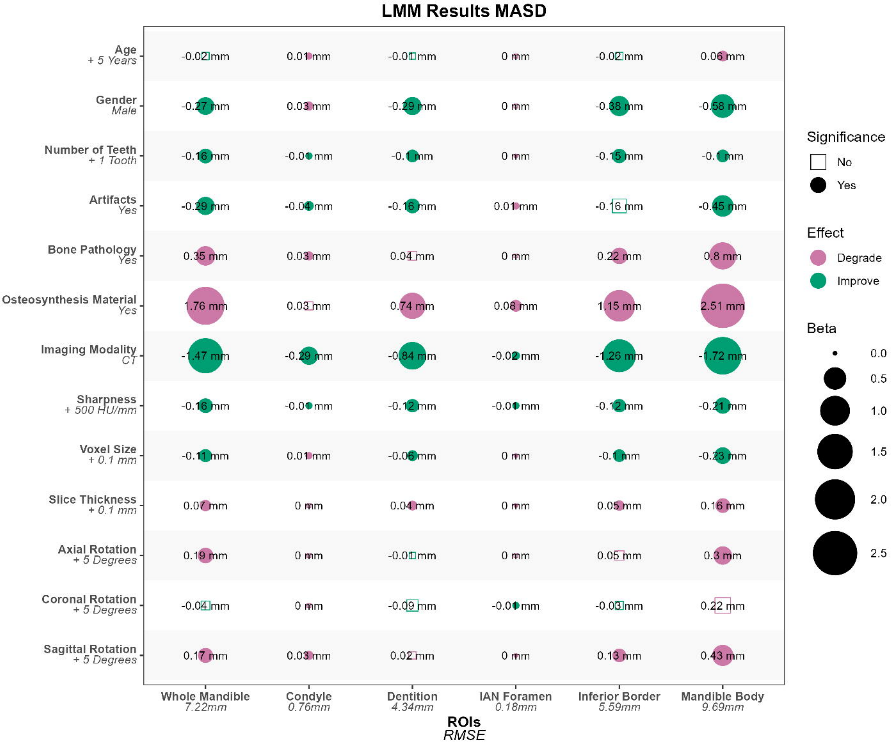
a
b
c
d
e
f
g
h
DSC
1
0.95
0.9
0.85
0.8
0.75
0.50
0.25
0
f
m
Gender
20
40
60
80
Age
500
750
1000
1250
150
HU Intensity
none
partial
full
Dentition
0
5
10
15
Number of Teeth
no
yes
Metal Artifacts
Bone Pathology
Osteosynthesis

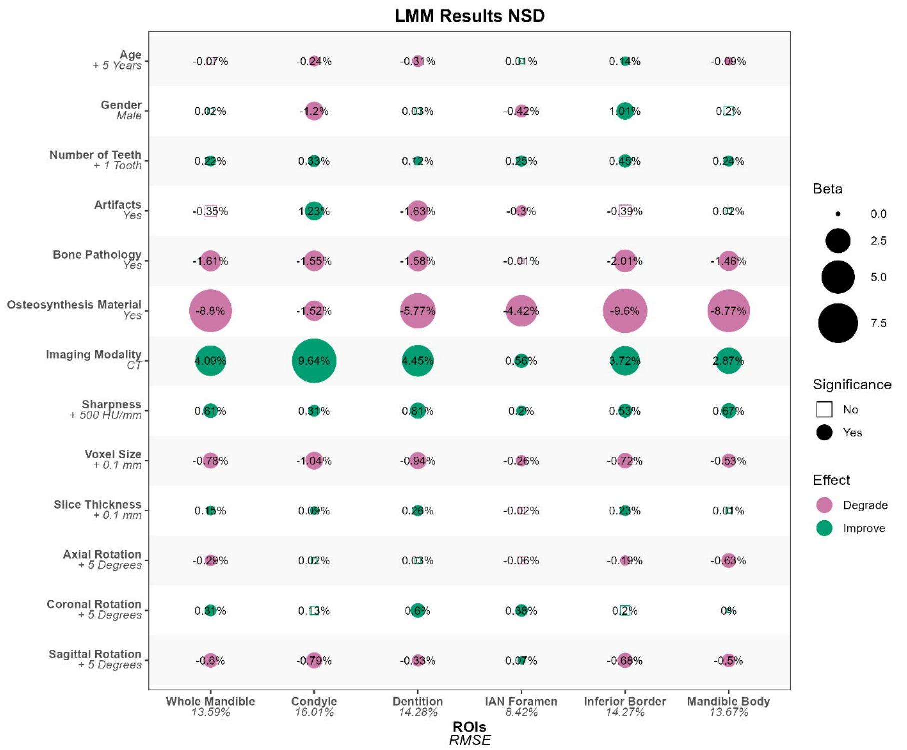

CBCT 1
CBCT 2
CBCT 3
CBCT 4
CBCT 5
CBCT 6
CBCT 7
CBCT 8
CBCT 9
CBCT 10
CBCT 11
CBCT 12
CBCT 13
CBCT 14
CBCT 15
CBCT 16
CBCT 17
CBCT 18
CBCT 19
CBCT 20
CBCT 21
CBCT 22
CBCT 23
CBCT 24
CBCT 25
CBCT 26
CBCT 27
CBCT 28
CBCT 29
CBCT 30
CBCT 31
CBCT 32
CBCT 33
CBCT 34
CBCT 35
CBCT 36
CBCT 37
CBCT 38
CBCT 39
CBCT 40
CBCT 41
CBCT 42
CBCT 43
CBCT 44
CBCT 45
CBCT 46
CBCT 47
CBCT 48
CBCT 49
CBCT 50
0 mm
1 mm
3 mm
5 mm
Distance (mm)

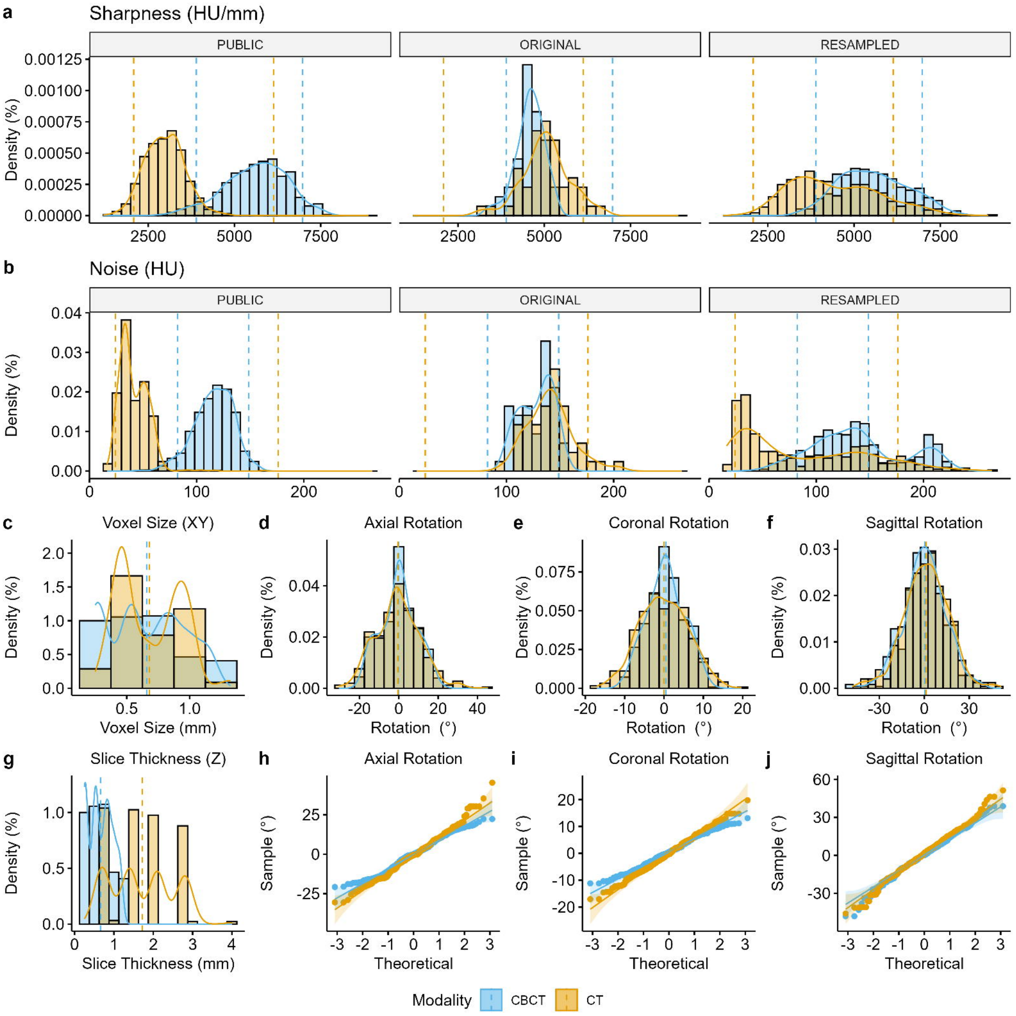

CT 1
CT 2
CT 3
CT 4
CT 5
CT 6
CT 7
CT 8
CT 9
CT 10
CT 11
CT 12
CT 13
CT 14
CT 15
CT 16
CT 17
CT 18
CT 19
CT 20
CT 21
CT 22
CT 23
CT 24
CT 25
CT 26
CT 27
CT 28
CT 29
CT 30
CT 31
CT 32
CT 33
CT 34
CT 35
CT 36
CT 37
CT 38
CT 39
CT 40
CT 41
CT 42
CT 43
CT 44
CT 45
CT 46
CT 47
CT 48
CT 49
CT 50
0 mm
1 mm
3 mm
5 mm
Distance (mm)


LMM Results DSC
Age
+ 5 Years
Gender
Male
Number of Teeth
+ 1 Tooth
Artifacts
Yes
Bone Pathology
Yes
Osteosynthesis Material
Yes
Imaging Modality
CT
Sharpness
+ 500 HU/mm
Voxel Size
+ 0.1 mm
Slice Thickness
+ 0.1 mm
Axial Rotation
+ 5 Degrees
Coronal Rotation
+ 5 Degrees
Sagittal Rotation
+ 5 Degrees
-0.06% -0.09% -0.29% -0.21% 0.13% -0.16%
0.99% -0.46% 0.76% 0.25% 1.47% 1.08%
0.38% 0.48% 0.2% 0.45% 0.6% 0.2%
-0.21% 0.98% -1.68% 0.16% -0.07% 0.03%
-0.71% -0.77% -0.39% -0.02% -1% -1.27%
-7.9% -1.81% -6.46% -7.7% -8.37% -8.23%
4.13% 8.35% 2.96% -0.4% 2.87% 3.12%
0.5% 0.29% 0.82% 0.5% 0.48% 0.61%
-0.16% -0.38% -0.34% -0.05% -0.11% -0.07%
-0.1% -0.17% -0.1% -0.09% -0.05% -0.15%
-0.51% -0.27% -0.14% -0.38% -0.33% -0.87%
-0.1% -0.36% 0.4% -0.28% -0.04% -0.57%
-0.69% -0.86% -0.26% 0.26% -0.62% -0.7%
Whole Mandible
13.52%
Condyle
16.94%
Dentition
14.58%
IAN Foramen
12.36%
Inferior Border
14.23%
Mandible Body
13.42%
ROIs
RMSE
Beta
2
4
6
8
Significance
No
Yes
Effect
Degrade
Improve


LMM Results HD95
Age
+ 5 Years
-0.08 mm
0.03 mm
-0.07 mm
0.01 mm
0.01 mm
0.61 mm
Gender
Male
-1.46 mm
0.13 mm
-1.14 mm
-0.02 mm
-2.1 mm
-2.79 mm
Number of Teeth
+ 1 Tooth
-0.82 mm
-0.04 mm
-0.49 mm
-0.02 mm
-0.67 mm
-0.16 mm
Artifacts
Yes
-1.09 mm
-0.22 mm
-0.88 mm
0.05 mm
-1.14 mm
-1.64 mm
Bone Pathology
Yes
2.04 mm
0.21 mm
0.44 mm
0 mm
1.7 mm
3.47 mm
Osteosynthesis Material
Yes
11.01 mm
0.02 mm
3.57 mm
0.34 mm
7.84 mm
16.17 mm
Imaging Modality
CT
-5.46 mm
-1.37 mm
-3.9 mm
-0.1 mm
-6.51 mm
-5.32 mm
Sharpness
+ 500 HU/mm
-0.7 mm
-0.04 mm
-0.51 mm
-0.01 mm
-0.5 mm
-0.63 mm
Voxel Size
+ 0.1 mm
-0.71 mm
0.01 mm
-0.33 mm
0.03 mm
-0.5 mm
-1.35 mm
Slice Thickness
+ 0.1 mm
0.42 mm
0.05 mm
0.25 mm
0.02 mm
0.32 mm
0.84 mm
Axial Rotation
+ 5 Degrees
0.97 mm
0.04 mm
0.18 mm
0 mm
0.43 mm
1.1 mm
Coronal Rotation
+ 5 Degrees
0.09 mm
0.02 mm
-0.24 mm
-0.03 mm
0.09 mm
2.58 mm
Sagittal Rotation
+ 5 Degrees
0.87 mm
0.17 mm
0.16 mm
-0.01 mm
0.86 mm
2.46 mm
Whole Mandible
30.32mm
Condyle
3.33mm
Dentition
18.35mm
IAN Foramen
0.76mm
Inferior Border
24.47mm
Mandible Body
38.85mm
ROIs
RMSE
Beta
0
4
8
12
16
Significance
No
Yes
Effect
Degrade
Improve


LMM Results MASD
Age
+ 5 Years
Gender
Male
Number of Teeth
+ 1 Tooth
Artifacts
Yes
Bone Pathology
Yes
Osteosynthesis Material
Yes
Imaging Modality
CT
Sharpness
+ 500 HU/mm
Voxel Size
+ 0.1 mm
Slice Thickness
+ 0.1 mm
Axial Rotation
+ 5 Degrees
Coronal Rotation
+ 5 Degrees
Sagittal Rotation
+ 5 Degrees
-0.02 mm
0.01 mm
-0.01 mm
0 mm
-0.02 mm
0.06 mm
-0.27 mm
0.03 mm
-0.29 mm
0 mm
-0.38 mm
-0.58 mm
-0.16 mm
-0.01 mm
-0.1 mm
0 mm
-0.15 mm
-0.1 mm
-0.29 mm
-0.04 mm
-0.16 mm
0.01 mm
-0.16 mm
-0.45 mm
0.35 mm
0.03 mm
0.04 mm
0 mm
0.22 mm
0.8 mm
1.76 mm
0.03 mm
0.74 mm
0.08 mm
1.15 mm
2.51 mm
-1.47 mm
-0.29 mm
-0.84 mm
-0.02 mm
-1.26 mm
-1.72 mm
-0.16 mm
-0.01 mm
-0.12 mm
-0.01 mm
-0.12 mm
-0.21 mm
-0.11 mm
0.01 mm
-0.06 mm
0 mm
-0.1 mm
-0.23 mm
0.07 mm
0 mm
0.04 mm
0 mm
0.05 mm
0.16 mm
0.19 mm
0 mm
-0.01 mm
0 mm
0.05 mm
0.3 mm
-0.04 mm
0 mm
-0.09 mm
-0.01 mm
-0.03 mm
0.22 mm
0.17 mm
0.03 mm
0.02 mm
0 mm
0.13 mm
0.43 mm
Whole Mandible
7.22mm
Condyle
0.76mm
Dentition
4.34mm
IAN Foramen
0.18mm
Inferior Border
5.59mm
Mandible Body
9.69mm
ROIs
RMSE
Significance
No
Yes
Effect
Degrade
Improve
Beta
0.0
0.5
1.0
1.5
2.0
2.5


LMM Results NSD
Age
+ 5 Years
Gender
Male
Number of Teeth
+ 1 Tooth
Artifacts
Yes
Bone Pathology
Yes
Osteosynthesis Material
Yes
Imaging Modality
CT
Sharpness
+ 500 HU/mm
Voxel Size
+ 0.1 mm
Slice Thickness
+ 0.1 mm
Axial Rotation
+ 5 Degrees
Coronal Rotation
+ 5 Degrees
Sagittal Rotation
+ 5 Degrees
-0.07% -0.24% -0.31% 0.01% 0.14% -0.09%
0.02% -1.2% 0.03% -0.42% 1.01% 0.2%
0.22% 0.33% 0.12% 0.25% 0.45% 0.24%
-0.35% 1.23% -1.63% -0.3% -0.39% 0.02%
-1.61% -1.55% -1.58% -0.01% -2.01% -1.46%
-8.8% -1.52% -5.77% -4.42% -9.6% -8.77%
4.09% 9.64% 4.45% 0.56% 3.72% 2.87%
0.61% 0.31% 0.81% 0.2% 0.53% 0.67%
-0.78% -1.04% -0.94% -0.26% -0.72% -0.53%
0.15% 0.09% 0.26% -0.02% 0.23% 0.01%
-0.29% 0.02% 0.03% -0.06% -0.19% -0.63%
0.31% 0.13% 0.6% 0.38% 0.2% 0%
-0.6% -0.79% -0.33% 0.07% -0.68% -0.5%
Whole Mandible
13.59%
Condyle
16.01%
Dentition
14.28%
IAN Foramen
8.42%
Inferior Border
14.27%
Mandible Body
13.67%
ROIs
RMSE
Beta
0.0
2.5
5.0
7.5
Significance
No
Yes
Effect
Degrade
Improve


a
Sharpness (HU/mm)
PUBLIC
ORIGINAL
RESAMPLED
Density (%)
b
Noise (HU)
c
Voxel Size (XY)
Voxel Size (mm)
d
Axial Rotation
e
Coronal Rotation
f
Sagittal Rotation
Rotation (°)
g
Slice Thickness (Z)
Slice Thickness (mm)
h
Axial Rotation
i
Coronal Rotation
j
Sagittal Rotation
Sample (°)
Theoretical
Modality
CBCT
CT