Accuracy and Precision of Mandible Segmentation and Its Clinical Implications

Gruber, Lennart Johannes; Egger, Jan; Bönsch, Andrea; Kraeima, Joep; Ulbrich, Max; van den Bosch, Vincent; Motmaen, Ila; Wilpert, Caroline; Ooms, Mark; Isfort, Peter

Published in:
Expert systems with applications

DOI:
10.1016/j.eswa.2023.122275

Publication date:
2024

Citation for published version (APA):
Gruber, L. J., Egger, J., Bönsch, A., Kraeima, J., Ulbrich, M., van den Bosch, V., Motmaen, I., Wilpert, C., Ooms, M., Isfort, P., Hölzle, F., & Puladi, B. (2024). Accuracy and Precision of Mandible Segmentation and Its Clinical Implications: Virtual Reality, Desktop Screen and Artificial Intelligence. *Expert systems with applications*, *239*, Article 122275. https://doi.org/10.1016/j.eswa.2023.122275

Expert Systems With Applications 239 (2024) 122275

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Accuracy and Precision of Mandible Segmentation and Its Clinical Implications: Virtual Reality, Desktop Screen and Artificial Intelligence

Lennart Johannes Gruber [a], Jan Egger [b,c,d], Andrea Bönsch [e], Joep Kraeima [f], Max Ulbrich [a], Vincent van den Bosch [g], Ila Motmaen [a], Caroline Wilpert [g,h], Mark Ooms [a], Peter Isfort [g], Frank Hölzle [a], Behrus Puladi [a,i,*]

[a] *Department of Oral and Maxillofacial Surgery, University Hospital RWTH Aachen, 52074 Aachen, Germany*
[b] *Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), 45131 Essen, Germany*
[c] *Center for Virtual and Extended Reality in Medicine (ZvRM), University Hospital Essen (AöR), 45147 Essen, Germany*
[d] *Institute of Computer Graphics and Vision, Graz University of Technology, 8010 Graz, Austria*
[e] *Visual Computing Institute, Faculty of Mathematics, Computer Science and Natural Sciences, RWTH Aachen University, 52056 Aachen, Germany*
[f] *Department of Oral and Maxillofacial Surgery, University of Groningen, UMCG Groningen, 9713 GZ Groningen, the Netherlands*
[g] *Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, 52074 Aachen, Germany*
[h] *Department of Diagnostic and Interventional Radiology, Medical Center, University of Freiburg, Faculty of Medicine, 79106 Freiburg, Germany*
[i] *Institute of Medical Informatics, University Hospital RWTH Aachen, 52074 Aachen, Germany*

ARTICLE INFO

Keywords:
Segmentation
Computer-assisted surgery
Virtual reality
Artificial intelligence
Oral and maxillofacial surgery

ABSTRACT

Objective: 3D modeling is a major challenge in computer-assisted surgery (CAS). Manual segmentation, as the gold standard, is tedious, time consuming, and particularly challenging for the mandible, while artificial intelligence (AI)-based segmentation is a promising and time-saving alternative. However, little is known about the clinical implications of various segmentation methods.

Method: In this cross-over study, ten mandibles were segmented in virtual reality (VR), on a desktop screen (DS) by five experts and via five AI models. The exported mandible models were evaluated using metrics, a public reference (PUB_{DS}), and blinded assessments by two radiologists.

Results: Average segmentation-to-volume accuracy (1 = poor, 5 = perfect) was comparable for human segmentation (VR: 4.56; DS: 4.33; PUB_{DS}: 4.55) and significant better than AI-based segmentation (AI: 3.80), while the average segmentation-to-segmentation accuracy revealed that DS (91.4 %/0.37 mm [Dice coefficient/average Hausdorff distance]) was more comparable to PUB_{DS} than to VR (90.1 %/0.44 mm). The precision of VR (96.8 %/0.14 mm) and DS (96.6 %/0.15 mm) was superior to PUB_{DS} (94.1 %/0.21 mm) and the AI method (89.2 %/0.60 mm). While VR was significantly faster than DS and PUB_{DS} for the manual segmentation methods ($p = 0.007/< 0.001$), in contrast, the AI method is not time sensitive due to its possible hardware scalability.

Conclusion: Accuracy and precision of mandible segmentation depends primarily on CT quality and anatomical site, which should be considered in clinical applications and the generation of AI training data and could negatively impact CAS. Although current AI models have perfect intra-model reliability, they demonstrate higher inter-model variability and are accompanied by invalid outliers making human review still necessary. In summary, the use of VR in manual segmentation showed high accuracy and precision overall while saving time, making it the preferred method over DS due to its good usability.

Abbreviations: VR, virtual reality; DS, desktop screen; AI, artificial intelligence; PUB, public CT dataset published by Wallner et al. (Wallner, Mischak, & Egger, 2019); PUB_{DS}, desktop-based segmentation method used and published by Wallner et al. (Wallner, Mischak, & Egger, 2019); CAS, computer assisted surgery; DICE, Sørensen–Dice coefficient; AVD, average Hausdorff distance; HU, Hounsfield unit; CT, computer tomography; CBCT, cone beam computer tomography; PSMRP, patient-specific mandibular reconstruction plates; TMJ, temporomandibular joint; HMD, head mounted device; STL, standard tessellation language; WQHD, wide quad high definition.

* Corresponding author at: Department of Oral and Maxillofacial Surgery, University Hospital RWTH Aachen, Pauwelsstraße 30, 52074 Aachen, Germany.
E-mail address: bpuladi@ukaachen.de (B. Puladi).

https://doi.org/10.1016/j.eswa.2023.122275
Received 10 April 2023; Received in revised form 12 September 2023; Accepted 19 October 2023
Available online 30 October 2023

1. Introduction

Personalized computer-assisted surgery (CAS) in the field of oral and maxillofacial surgery has advanced in recent years with digital technologies and improvements in computed tomography (CT) imaging (Christensen, Weimer, Beaudreau, Rensberger, & Johnson; Minnema et al., 2022). The three main challenges of CAS are CT image reconstruction, segmentation, and surgical planning (Minnema et al., 2022). In particular, segmentation and surgical planning require intense human interaction by medical professionals (Nysjö, 2016; Zhao & Xie, 2013). Segmentation is the labeling of voxels in a 3D volumetric image to derive a 3D surface model (Bryan et al., 2014). All subsequent steps in personalized surgery are based on these 3D models. Therefore, any inaccuracy or imprecision in this process decreases the corresponding quality of CAS, including virtual surgical planning, patient-specific implants, cutting guides, the application of augmented reality and robot-guided surgery, and thereby the final surgical outcome (van Eijnatten et al., 2018).

Although manual segmentation on a 2D desktop screen (DS) is time consuming, tedious, and prone to interindividual differences, it is still considered the gold standard (Bryan et al., 2014; Minnema et al., 2022; Ulbrich et al., 2023). Segmentation of the mandible is more challenging due to structural variation, its complex morphology, and poor joint contours. It is further complicated by connections between the maxillary and mandibular teeth or the presence of artifacts resulting from prosthetics (Torosdagli et al., 2017). However, segmentation of the mandible is necessary for a wide range of CAS scenarios. This includes 3D printing of patient-specific drilling/cutting guides, implants for mandibular/temporomandibular joint surgery, orthognathic surgery, complex trauma reconstruction, aesthetic procedures and dental implantology (Greenberg, 2018).

Efforts have been made to automate the complete segmentation process, including advanced thresholding methods and statistical shape model methods (van Eijnatten et al., 2018). Recently, methods using artificial intelligence (AI), in particular conventional neural networks, have been increasing in popularity (Minnema et al., 2022). However, AI-based segmentation methods are highly dependent on training data, causing AI models to adopt various errors (Thambawita et al., 2022; Yu et al., 2020). Therefore, high-quality training data that need to be manually generated are required (Yu et al., 2020). Furthermore, the use of AI in the medical context is strictly regulated in the US and EU, respectively (Muehlematter, Daniore, & Vokinger, 2021; Pesapane, Volonté, Codari, & Sardanelli, 2018), with accountability being a concern. For example, an error in AI-based segmentation could lead to incorrect clinical decisions (Pesapane et al., 2018). Consequently, even with AI-based segmentation, medical professionals still need to scrutinize the results and make adjustments if necessary. Therefore, efforts should be made to improve the speed and reduce the inter-rater variability of manual segmentation.

Unlike manual DS segmentation, which offers no spatial depth perception and a very flat learning curve (Ulbrich et al., 2023), segmentation in virtual reality (VR) could enable a more efficient way to interact with 3D data (Nysjö, 2016). VR-based segmentation has already been shown to have significant speed advantages over a DS method for fibula and os coxae segmentation during a training program (Ulbrich et al., 2023). However, for the mandible, different manual segmentations have a strong influence on the accuracy of the 3D surface model (Engelbrecht, Fourie, Damstra, Gerrits, & Ren, 2013; Fourie, Damstra, Schepers, Gerrits, & Ren, 2012). Although computer-assisted mandibular reconstruction leads to a reduction in ischemic time, total operation time, reconstruction time, and length of hospital stay (Powcharoen, Yang, Yan Li, Zhu, & Su, 2019), the benefits of its improved accuracy have yet to be demonstrated due to a lack of uniformity in planning and evaluation methods (van Baar, Forouzanfar, Liberton, Winters, & Leusink, 2018).

Therefore, it is necessary to investigate the accuracy and precision of different methods and their clinical impacts in the context of CAS workflows. In this cross-over study, a set of mandibles from a publicly available CT dataset (Wallner, Mischak, & Egger, 2019) (PUB) was segmented by trained medical professionals using both segmentation methods (VR and DS) and subsequently compared to the results of the publicly available method (PUB_{DS}) that served as a reference. Additionally, the PUB data set has been segmented via AI segmentation models and was also compared with the segmentation methods VR, DS and PUB_{DS}. For all methods, the variability with respect to the method and different regions of the mandible was determined using quantitative and qualitative measurements.

2. Material and methods

2.1. Study

Ten edentulous mandibles were segmented by five trained medical professionals (P_1 to P_5, all male, aged 33.6 ± 0.89 years) in VR and DS by a cross-over allocation rule (Fig. 1). For this purpose, a public DICOM dataset featuring ten CT scans of the skull with edentulous mandibles was used (Wallner et al., 2019) (PUB), since this is the most common modality for generating 3D models of the mandible (Qiu et al., 2021). The medical professionals were in their second to fourth year of residency to obtain a dual degree (board-certified physicians and dentists) from the Department of Oral and Maxillofacial Surgery, RWTH Aachen University Hospital, and were trained in bone segmentation in VR and DS (Ulbrich et al., 2023). The Ethics Committee of the Medical Faculty of RWTH Aachen University approved our study (approval number EK 471/20).

Segmentation in VR was performed using Elucis (Realize Medical Inc., Ottawa, Canada) with a head-mounted display (HMD) (HTC Vive Pro), an HTC Vive Controller 2.0 (HTC Corporation, Taoyuan, Taiwan), and a VR stylus (Ink Pilot Edition, Logitech International SA, Apples, Switzerland) while participants sat at a table covered with a VR Ink Drawing Mat (A1 size) optimally positioned between two HTC Base Stations 2.0 for tracking (Fig. 2a). Segmentation on a DS using 3D Slicer (www.slicer.org, version 4.11.20.210226) involved participants sitting at a workstation in front of a 2D desktop screen using a computer mouse and keyboard (Fig. 2b). The two applications (VR and DS) were run on the same workstation. The duration of each segmentation session was recorded. In addition, the participants completed a Likert questionnaire after each case was segmented, and a final questionnaire was completed after participation. To evaluate the utility of the VR stylus, a final standardized questionnaire was administered.

Furthermore, five different AI segmentation models (AI_1 to AI_5) were used to perform the same task (Table 1). Four of the AI models were used by the authors/companies themselves, who kindly provided us with the segmented models (see acknowledgements). The public AI model was implemented using the following instructions (https://github.com/Maxlo24/AMASSS_CBCT) to create the segmented models.

2.2. Evaluation

For evaluation, the label maps in VR, DS and AI were exported. In addition, the segmentation results were obtained from Wallner et al. (Wallner et al., 2019), where two medical professionals (W_A and W_B) segmented all models slice by slice on a 2D desktop screen using a contour path (PUB_{DS}) (Wallner et al., 2019). While the label maps from DS could be directly exported (3D Slicer), a 3D representation in VR (Elucis), and the Contour Segmentation Object (CSO) models of the PUB dataset (MeVisLab CSO) had to be converted into label maps using MeVisLab (version 3.6.1.9) before. For one of the AI models only STL files were provided, which had to be converted into label maps.

Afterwards, a group-wise comparison was performed to determine the precision of the methods. Additionally, the VR, DS and AI results were compared to the PUB_{DS} segmentations as a reference for

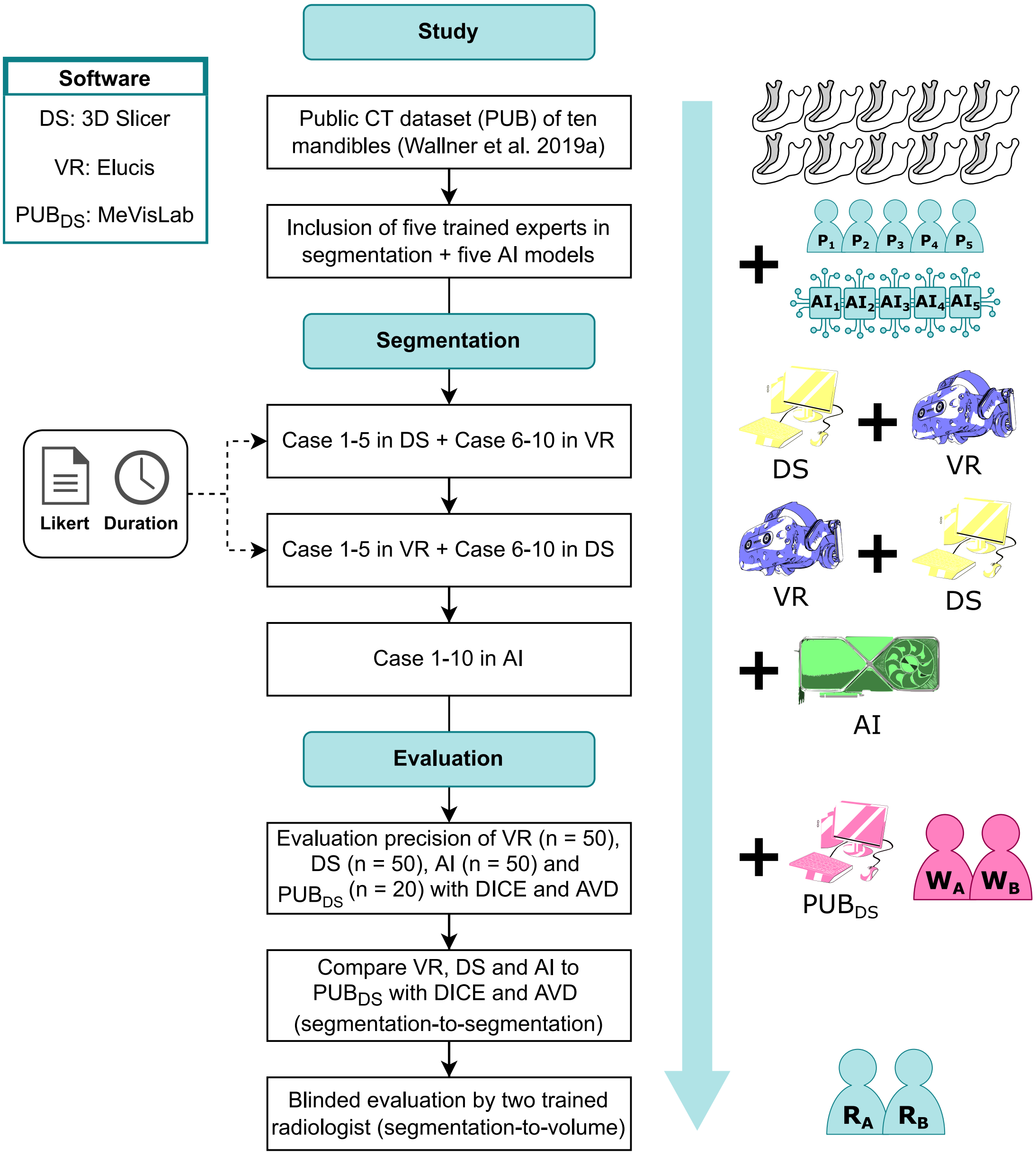
Study
Software
DS: 3D Slicer
VR: Elucis
PUB$_{DS}$: MeVisLab
Public CT dataset (PUB) of ten mandibles (Wallner et al. 2019a)
Inclusion of five trained experts in segmentation + five AI models
P1 P2 P3 P4 P5
AI1 AI2 AI3 AI4 AI5
Segmentation
Likert
Duration
Case 1-5 in DS + Case 6-10 in VR
DS
VR
Case 1-5 in VR + Case 6-10 in DS
VR
DS
Case 1-10 in AI
AI
Evaluation
Evaluation precision of VR (n = 50), DS (n = 50), AI (n = 50) and PUB$_{DS}$ (n = 20) with DICE and AVD
PUB$_{DS}$
W$_A$ W$_B$
Compare VR, DS and AI to PUB$_{DS}$ with DICE and AVD (segmentation-to-segmentation)
Blinded evaluation by two trained radiologist (segmentation-to-volume)
R$_A$ R$_B$

Fig. 1. Study design featuring study preparation, segmentation, and evaluation steps. The segmentation of a public CT dataset featuring 10 skulls with edentulous mandibles (PUB) was manually performed in two different working environments: on a desktop screen (DS) and in virtual reality (VR). Additionally the PUB dataset was segmented via 5 AI-based segmentation models (AI$_1$-AI$_5$). The public segmentation method (PUB$_{DS}$) was used for comparison. To evaluate precision and allow a comparison with the public dataset, the Dice coefficient (DICE) and average Hausdorff distance (AVD) were calculated. Accuracy was assessed through a segmentation-to-segmentation comparison (PUB$_{DS}$ as a reference) and segmentation-to-volume comparison (subjective evaluation) by two independent radiologists who scored performance (1 = poor, 5 = perfect) in a blinded fashion.

segmentation-to-segmentation accuracy. The precision heatmaps were then generated for each method (VR, DS and AI; $\boldsymbol{M}$) given $\bigcup_{i=0}^{n}\boldsymbol{M}_i - \bigcap_{i=0}^{n}\boldsymbol{M}_i$ and the accuracy heatmaps according to the external reference ($\boldsymbol{P}$) given $\bigcup_{i=0}^{n}\boldsymbol{M}_i - \bigcap_{i=0}^{n}\boldsymbol{P}_i$. All were visualized on a scale between 0 mm (blue) and 3 mm (red). The Sørensen–Dice coefficient (DICE) and average Hausdorff distance (AVD) were calculated using a Python script and the SlicerRT comparison module in 3D Slicer.

In addition, a blinded qualitative evaluation was carried out by two independent radiologists with professional experience (R$_A$ and R$_B$; final-year residents just before specialist examination) to determine segmentation-to-volume accuracy. Accuracy was graded on a Likert scale (1 = poor to 5 = perfect). Thus, the whole mandible and five regions of special clinical interest (alveolar crest, capitulum, foramen mandibulae, lower edge, and the outer surface) were scored.

2.3. *Statistical analysis*

The R programming language was used for statistical analysis. A p-level < 0.05 was considered significant. Analysis of variance (ANOVA) was used for normally distributed data and Kruskal-Wallis test was used for non-normally distributed data. Normal distribution was examined using the Shapiro–Wilk test. When post-hoc analysis was required, the Tukey test was used for normally distributed data and the Mann–Whitney *U* test was used for non-normally distributed data. P values were corrected for multiple testing using the Bonferroni method.

3. Results

The study resulted in 150 segmented models (50 in VR, 50 in DS and

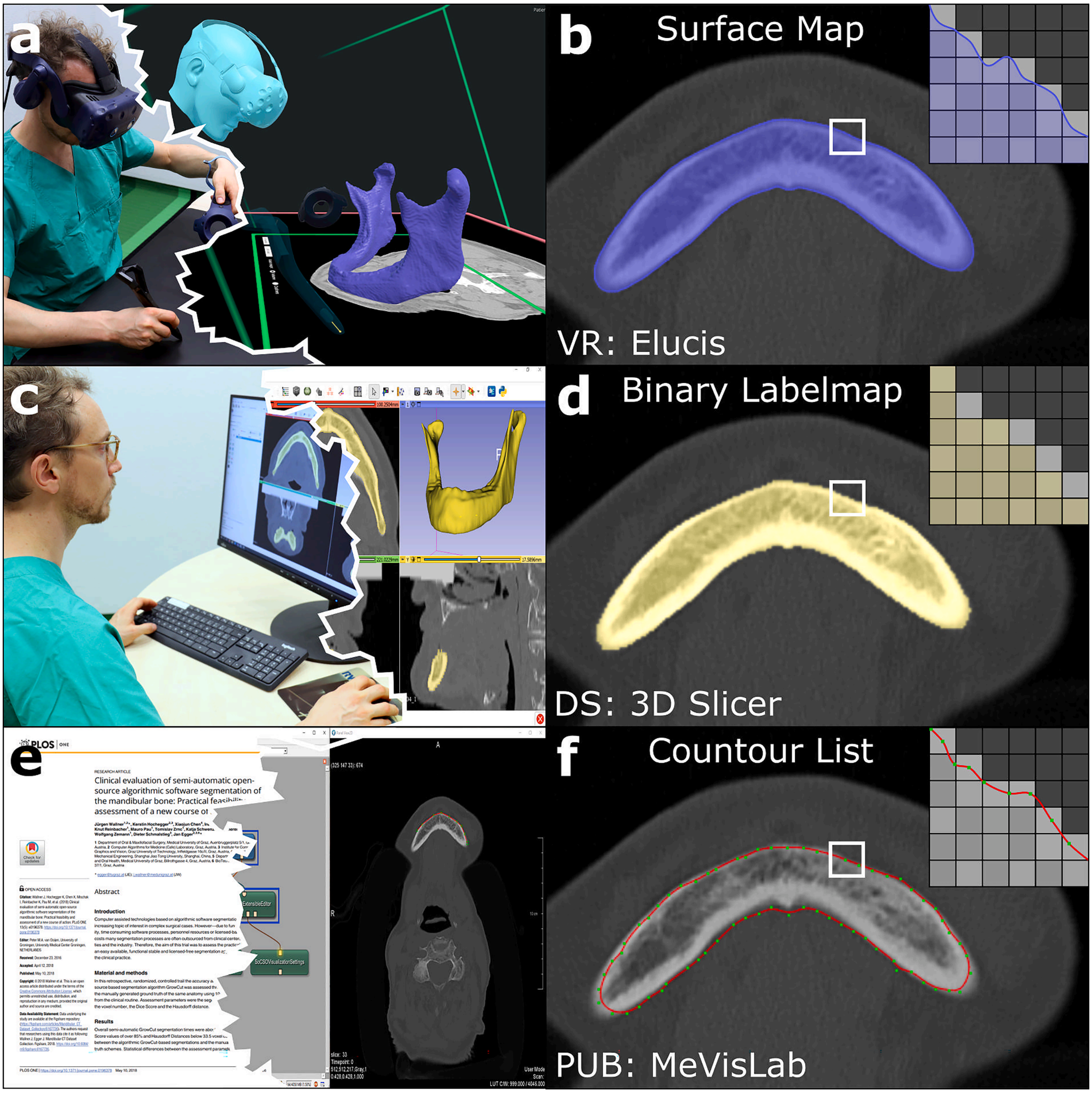
a
b
Surface Map
VR: Elucis
c
d
Binary Labelmap
DS: 3D Slicer
e
PLOS | ONE
Clinical evaluation of semi-automatic open-source algorithmic software segmentation of the mandibular bone: Practical feasibility assessment of a new course of
Abstract
Introduction
Material and methods
Results
f
Countour List
PUB: MeVisLab

Fig. 2. (**a** + **b**) Technical study setting: Segmentation in virtual reality (VR) with the creation of a 3D surface map where the surface boundary passes through individual voxels (software: Elucis); (**c** + **d**) Segmentation on a desktop screen (DS) with the creation of a binary label map where the surface boundary passes between individual voxels. A voxel is completely included or excluded in the model (software: 3D Slicer); (**e** + **f**) Public segmentation published by Wallner and colleagues 2019 (PUB_{DS}), also on a desktop screen with the creation of a contour path. After setting individual points, they were connected by a line, whereby the surface boundary also ran through individual voxels on each slice (software: MeVisLab).

50 in AI). In addition, the 20 publicly available (PUB_{DS}) models (from two subjects) were included for comparison. Qualitative assessments of segmentation-to-volume accuracy (1 = poor, 5 = perfect) by radiologists showed that the manual methods were not associated with significant performance differences, with VR having a score of 4.56 ± 0.45 (mean ± SD), DS 4.33 ± 0.59, and PUB_{DS} 4.55 ± 0.43 (Kruskal–Wallis test, $p = 0.11$). While including the AI-based method (AI 3.80 ± 0.86) led to significant differences (Kruskal–Wallis test, $p < 0.001$). In the pairwise comparison, the AI models performed worse than the other methods (VR, $p = 0.002$; DS $p < 0.001$; PUB_{DS}, $p < 0.001$). Among the manual methods, the VR method was faster (15.9 ± 9.0 min) compared to the DS method (21 ± 10.4 min) both being significantly faster than the PUB_{DS} method (40.7 ± 4.5 min). The segmentation time required by the five different AI models was not recorded as it is completely hardware dependent. Although the VR (35.7 ± 11.2 ml), DS (33.6 ± 10.5 ml) and

Table 1
Applied AI models.

Related Publication	Availability	Modality	Cases / Centers	Country
Gillot et al., 2022	Public	CBCT	618 / 7	Multicenter
Ileşan, Beyer, Kunz, & Thieringer, 2023	In-house	CT	160 / 1	Basel, Switzerland
Verhelst et al., 2021	Commercial	CT/ CBCT	NA†	Leuven, Belgium
Pankert et al., 2023	In-house	CT	307 / 1	Aachen, Germany
Xu et al., 2021	In-house	CT	230 / 1	Shanghai, China

† Due to business secret.

AI (32.5 ± 10.7 ml) models tended to have larger volumes than the PUB$_{DS}$ models (31.3 ± 10.2 ml). This trend showed no significance (ANOVA, p = 0.36) (Fig. 3; Table 2).

A more detailed analysis revealed very high precision overall associated with both VR (DICE of 96.8 ± 2.3 % and an AVD of 0.149 ± 0.108 mm) and DS (DICE of 96.6 ± 2.6 % and an AVD of 0.154 ± 0.113 mm). Despite fewer raters, precision was poorer for PUB$_{DS}$, with a DICE of 94.1 ± 1.1 % and an AVD of 0.211 ± 0.039 mm. The AI method showed the worst DICE with 89.2 ± 4.3 % and an AVD of 0.603 ± 0.511 mm compared to the other methods (Fig. 4; Supplementary Fig. 1; Table 2). Regardless of the method, variations seemed to be dependent on the anatomical site of the mandible. Sites that showed a large difference were the alveolar crest, capitulum, mandibular foramen, and lower edge. However, the outer surfaces of the mandible were very homogeneous. In cases of the AI-based method, abrupt interruptions of the continuity of the mandibular corpus occurred (Fig. 5). Nevertheless, irrespective of the method and anatomical site, the most influential factor affecting homogeneity was CT quality, which depends on the voxel dimensions, the reconstruction kernel, and the presence of artifacts (e.g., patient motion during the CT scan) (Supplementary Table 1; Fig. 6; Supplementary Fig. 2; Supplementary Table 2).

When PUB$_{DS}$ acted as the reference in the segmentation-to-segmentation comparison, DICE and AVD were 90.1 ± 2.2 % and 0.443 ± 0.096 mm for VR, 91.4 + 2.1 % and 0.373 ± 0.110 mm for DS and 88.8 ± 3.9 % and 0.544 ± 0.422 mm for AI respectively, making them significantly different (ANOVA, p < 0.001; Fig. 4c,d and 6). However, there were only notable outliers in the AI method caused by partially broken mandibular continuity (Fig. 5). In detail, the alveolar crest and capitulum appeared less inhomogeneous in comparison (Fig. 7; Supplementary Table 3). However, for the different anatomical sites, the various methods showed very different ranges in the qualitative evaluation. In particular, the capitulum and foramen mandibulae varied (Supplementary Fig. 1a). When looking at the different methods, the performance was good to perfect (good = 4, perfect = 5) overall, but with different variances, while the AI-based methods had some outliers (Fig. 8; Supplementary Table 3).

Table 2
Accuracy (segmentation-to-volume) assessed by two radiologists, accuracy (segmentation-to-segmentation) and precision in DICE (%) and AVD (mm), as well as speed (min) and volume (ml) compared between different methods (VR, DS, AI, PUB$_{DS}$).

	Mean (SD)	VR	DS	AI	PUB$_{DS}$	*P*
Accuracy (Segmentation-to-Volume)	Overall Rating	4.6 ± 0.4	4.3 ± 0.6	3.8 ± 0.9	4.6 ± 0.4	< 0.001
Accuracy (Segmentation-to-Segmentation)	DICE (%)	90.1 ± 2.2	91.4 ± 2.1	88.8 ± 3.9	NA[†]	< 0.001
	AVD (mm)	0.443 ± 0.096	0.373 ± 0.110	0.544 ± 0.422	NA[†]	< 0.001
Precision	DICE (%)	96.8 ± 2.3	96.6 ± 2.6	89.2 ± 4.3	94.1 ± 1.1	< 0.001
	AVD (mm)	0.149 ± 0.108	0.154 ± 0.113	0.603 ± 0.511	0.211 ± 0.039	< 0.001
Speed	min	15.9 ± 9.0	21.0 ± 10.4	NA*	40.7 ± 4.5	< 0.001
Volume	ml	35.7 ± 11.2	33.6 ± 10.5	32.5 ± 10.7	31.3 ± 10.2	0.36

[†] Serves as ground truth and is therefore not available. *Has not been recorded because the speed of AI models is completely hardware dependent.

However, the Likert-type questions (scored from 1 to 7) for the manual segmentation methods, as described in Table 3, showed that VR was clearly superior to DS in all items queried. Segmentation in VR was described as easier to reproduce (5.8 vs. 4.6) and generally easier to perform (6.2 vs. 4.4). In addition, subjects perceived segmentations in VR to be more appropriate for a CAS (6.1 vs. 4.9) and more compatible with the clinical routine (6.0 vs. 4.4). VR was rated better than DS when participants were asked to perceive the anatomical structures of the mandible (6.4 vs. 4.5). The temporomandibular joint, a particularly difficult region to segment, also appeared to be easier to separate from the skull in VR (5.8 vs. 3.9) than in DS, as is filling cavities to create a solid model (6.0 vs. 4.5). The ability to concentrate and work accurately was also rated higher in VR than in DS (6.0 vs. 4.8 and 6.0 vs. 4.7, respectively).

In response to the final questions, all medical professionals preferred VR as a working environment for segmentation. However, three of five stated that they had a better visual representation of the individual slices (axial, sagittal, and coronal) in the DS working environment. Four experts preferred segmentation with the VR stylus instead of the standard HTC Vive controller in the VR working environment. This is consistent with the finding that segmentation with the VR stylus was perceived as good (6.2 ± 0.4). In contrast, the accuracy of the VR stylus was rated

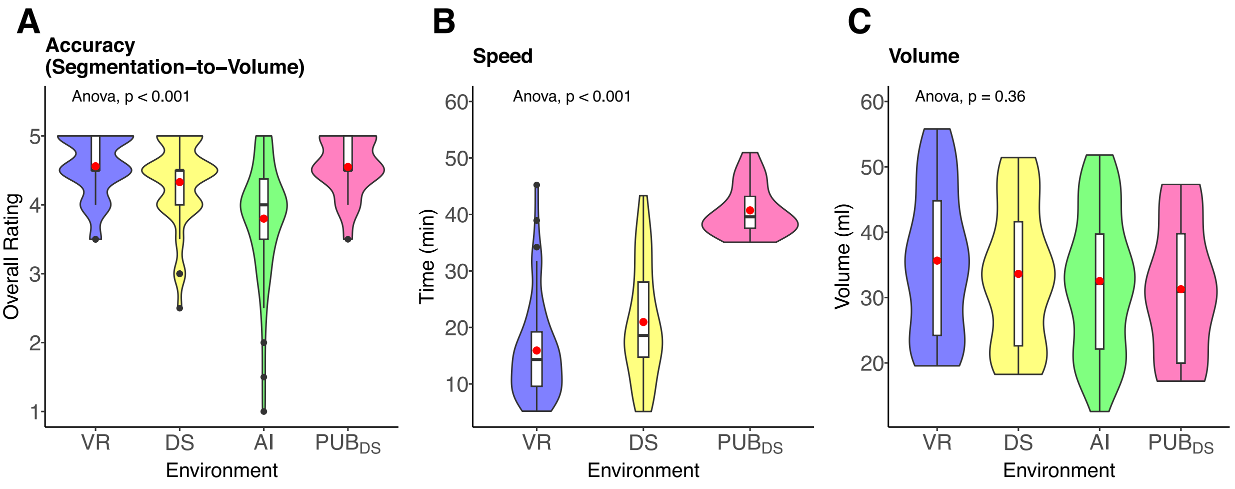

Fig. 3. Comparison of the (**a**) overall accuracy by rating, (**b**) time required for segmentation, and (**c**) the created volumes of the segmentations. Virtual reality (VR) is represented in violet, desktop screen (DS) in yellow, artificial intelligence (AI) in green and the public segmentation dataset (PUB$_{DS}$) (Wallner et al., 2019) in pink. (**a**) The overall rating scores reflecting accuracy (segmentation to volume) by two independent radiologists (1 = poor, 5 = perfect) are shown on the y-axis, (**b**) time in minutes is shown on the y-axis, and (**c**) segmented volume in ml is shown on the y-axis. Segmentation method on the x-axis. The violin plots (colored) include a boxplot (white), with the mean value marked as a red point. Black points are outliers.

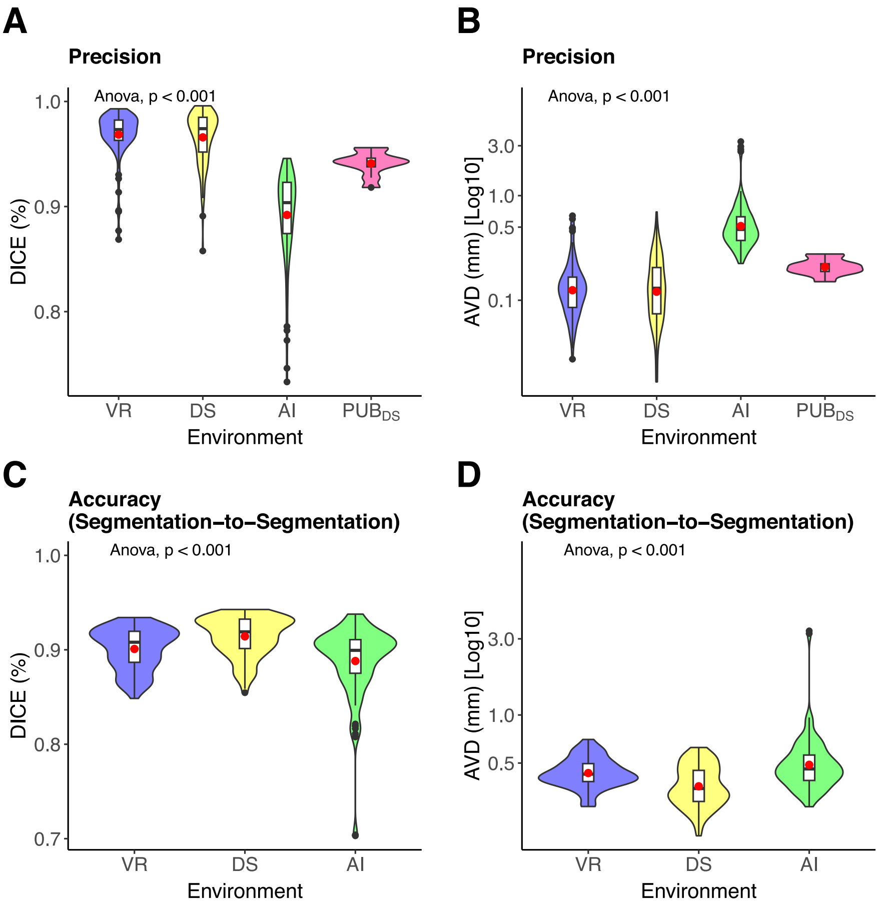

Fig. 4. **(a–b)** Comparison of the precision of the segmentations in virtual reality (VR) in violet, on a desktop screen (DS) in yellow, artificial intelligence (AI) in green and the public segmentation published by Wallner and colleagues (Wallner et al., 2019) in pink (PUB_{DS}). (**a**) The Dice coefficient (DICE) is shown on the y-axis, and (**b**) average Hausdorff distance (AVD) is shown on the y-axis (Log_{10} scale). (**c–d**) Accuracy (segmentation to segmentation) of VR and DS segmentation with PUB_{DS} segmentation as a reference. (**c**) DICE is shown on the y-axis, and (**d**) AVD is shown on the y-axis (Log_{10} scale). (**a–d**). The segmentation method is shown on the x-axis. The violin plots (colored) include a boxplot (white), with the mean value marked as a red point. Black points are outliers.

lower (4.8 ± 1.5). The precise grip of the VR stylus was perceived as positive (5.8 ± 1.5).

4. Discussion

Despite the advantages of speed, learnability, user satisfaction, and 3D visualization of a VR over a DS working environment, it was unclear how the different methods affected the overall and site-specific accuracy and precision of mandibular segmentation. As an increasing number of AI models have been developed with promising results for mandibular segmentation (Qiu et al., 2021), it has been unclear where they should be placed in terms of performance compared to the aforementioned manual segmentation methods.

Our study showed the following: First, the overall inter-rater precision of the DS and VR method was better compared to the PUB_{DS} and the AI method. Regarding the accuracy of segmentation-to-volume, we found no differences between manual segmentation methods. The segmentation-to-segmentation comparison showed that the DS method was more similar to the PUB_{DS} method than the VR method was to the DS method, but both reached sufficient DICE and AVD, which indicates a comparable use in clinical practice of VR, DS and PUB_{DS} (Wallner et al., 2018). In this regard, the AI method showed the worst performance and had noticeable outliers for both the segmentation-to-volume and segmentation-to-segmentation comparison.

At the same time and regardless of the method, the accuracy and precision were dependent on the anatomical location (alveolar crest, capitulum, foramen mandibulae, and lower edge), contrary to the external surfaces of the mandibles, which were very homogeneous. However, the most important factor affecting accuracy and precision seemed to be the underlying CT quality. Since the AI models (A_1-A_5) were trained on CTs/CBCTs of a specific quality and from specific vendors, this could be an explanation for the very low precision. Interestingly, the AI models that were partially or fully trained on CBCTs showed higher performance. One possible explanation could be the voxel size of CBCTs (Gaêta-Araujo et al., 2020), which is often set as smaller than that of CTs by default, potentially leading to better annotated training data. In this regard, manual segmentation, which showed variability depending on the anatomical region and is required for the development of AI models, could also be one reason for the analogous variability of the AI method. In one respect, however, the AI-based method will always be superior when it comes to the time required for segmentation, since the speed can be scaled by hardware or models can be automatically segmented overnight.

Nevertheless, among the manual segmentation methods, VR showed clear time advantages and was considered the preferred working environment. The PUB_{DS} method, which involved drawing a contour path to outline the corresponding bone layer by layer (Fig. 2e,f), was the most time-consuming method (Fig. 3b). Although the first thought might be that this is a more precise method for mandible segmentation, our results showed that VR, DS yielded a higher DICE, indicating higher precision (ANOVA and post-hoc Tukey test, p = 0.002 and p = 0.005, respectively). In this context, precision indicates how close

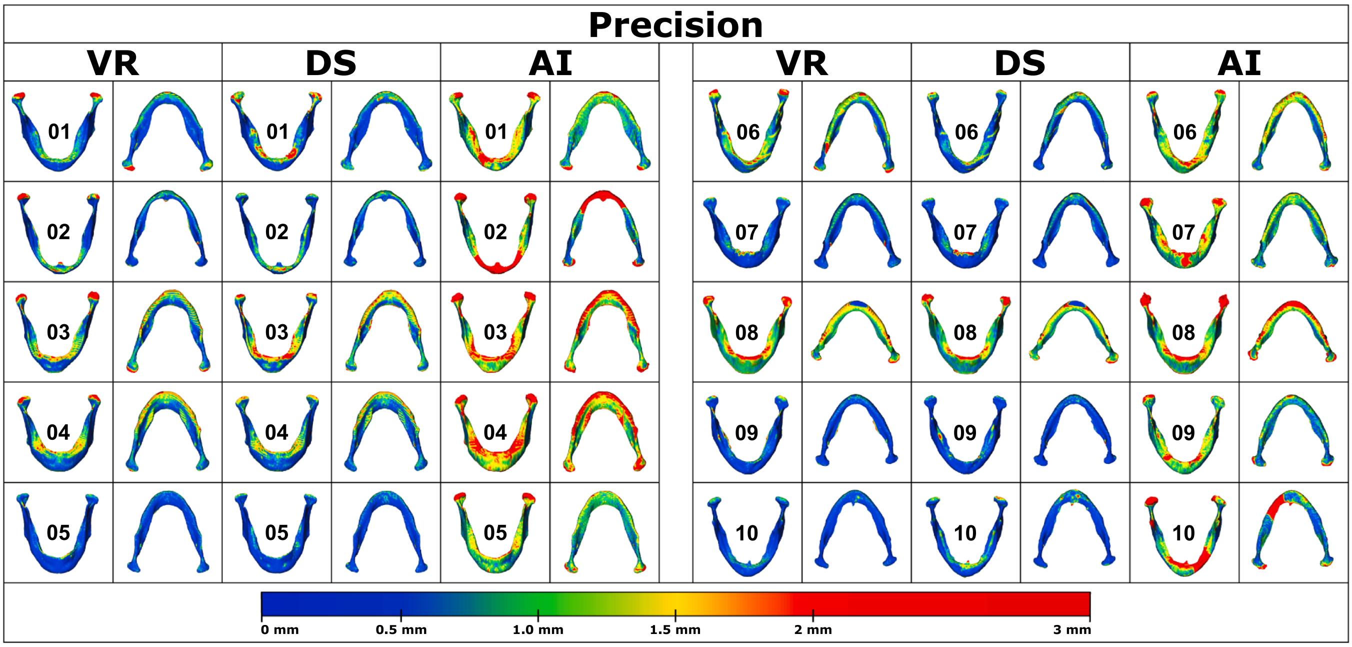

Fig. 5. Precision of the methods according to group-wise comparison of the segmentation: Virtual reality (VR), desktop screen (DS) and artificial intelligence (AI). The heat maps were made by overlaying the segmentations done by all five subjects and AI models per mandible, which show regions of high (red 2–3 mm), medium (green/yellow 0.5–2 mm), and low (blue 0–0.5 mm) variance.

segmentations are to each other (Hofer, Strauß, Koulechov, & Dietz, 2005). The reason why both VR (Elucis) and DS (3D Slicer) yielded a higher precision is likely because of the limitation in the investigators' (P_1 to P_5; all five medical professionals) degrees of freedom in interactions facilitated by the use of semi-automatic techniques. These include limiting the selection (i.e., segmentation) based on a Hounsfield unit (HU) threshold, logical operations to select/remove a set of selections, and filtering functions, such as closing to fill holes, which limit users' degrees of freedom. In this context, it is important to understand that there are different technical methods of segmentation. For example, 3D Slicer (DS) allows the performance of voxel-based segmentation using a binary label map, whereas in Elucis (VR), a 3D surface map is generated directly, allowing selection at the subvoxel level. Similarly, counter-selection in MeVisLab (PUB_{DS}) allows the definition of a path via points and is performed at the subvoxel level (Fig. 2). Therefore, the ability to work on the subvoxel level leads to a higher degree of freedom. In this context, AI has the advantage that it is independent of interactions and always leads to the same results with the same data input. However, it also shows the disadvantage that current AI models for mandibular segmentation are highly dependent on training data (Thambawita et al., 2022; Yu et al., 2020) and different AI models lead to very different results. Therefore, intra-model reliability will always be high, but inter-model reliability showed low precision.

However, precision is not equal to accuracy. Accuracy indicates how close a given set of segmentations is to its true value (Hofer et al., 2005). In that regard, it is important to note that CT image reconstructions themselves are not identical to the ground truth but are also an abstraction of patients' anatomy. In the past, studies were performed in which CT image reconstructions were compared with laser scans of cadaver bones, revealing an accuracy between 0.16 ± 0.06 and 0.38 ± 0.29 mm, depending on the region (Lalone, Willing, Shannon, King, & Johnson, 2015). In a study featuring mandibular cadavers, a comparison of segmentation in cone beam computed tomography (CBCT) and laser-scanned surfaces showed an accuracy of 0.330 ± 0.427 mm for experienced users and 0.763 ± 0.392 mm for inexperienced users (Fourie et al., 2012). However, obtaining a ground truth with a laser scan is not possible for living humans. Therefore, clinical studies on CAS have compared either preoperative and postoperative segmentation (i.e., 3D surface models), preoperative planning and postoperative segmentation, or preoperative and postoperative CT scans (van Baar et al., 2018). This issue is further complicated by a variety of possible evaluation metrics (Taha & Hanbury, 2015).

For our study, we used the overlap-based DICE and distance-based AVD metrics in accordance with recommendations for evaluating segmentation precision (Müller, Soto-Rey, & Kramer, 2022). PUB_{DS} was used as a possible reference in the segmentation-to-segmentation comparison (Müller et al., 2022). However, this should not be seen as a true ground truth, since the reference is also prone to errors. Therefore, in addition to a segmentation-to-segmentation comparison, a blinded segmentation-to-volume comparison was conducted by two radiologists. In this regard, subjective evaluation is the most common method (Gelasca, Ebrahimi, Farias, Carli, & Mitra, 2004; Wang, Wang, & Zhu, 2020).

In the segmentation-to-segmentation comparison, the DS method was more similar to the PUB_{DS} method, probably because it is more similar to a slice-by-slice approach, whereas the VR method focuses on spatial work. Both manual methods showed no significant differences in terms of segmentation-to-volume comparisons. In contrast, the AI method was only slightly worse than the manual segmentation methods in terms of accuracy, but had considerable outliers due to loss of continuity in the mandibular corpus. Nevertheless, regardless of the method there were significant differences between the individual cases (Fig. 6; Supplementary Fig. 2). The quality of the image reconstruction from the CT scan seemed to be the most important factor affecting accuracy and precision. Tube current (mA), tube voltage (kV), pitch, number of rotations, voxel size, slice thickness, and reconstruction filter are the main parameters in CT protocols (Willemink & Noël, 2019). Image reconstruction is performed using two basic methods: filtered back projection (FBP) and iterative reconstruction (IR), whereby the latter has become the industry standard in recent years (Minnema et al., 2022; Willemink & Noël, 2019).

The possible resolution of errors in 3D surface models are determined by the voxel size of the underlying volume during segmentation (Noser, Heldstab, Schmutz, & Kamer, 2011). For anisotropic voxel sizes (i.e., no equal spacing between the x, y, and z directions) (Supplementary Table 1), this has a particularly negative effect on surface divergence (Fig. 4,

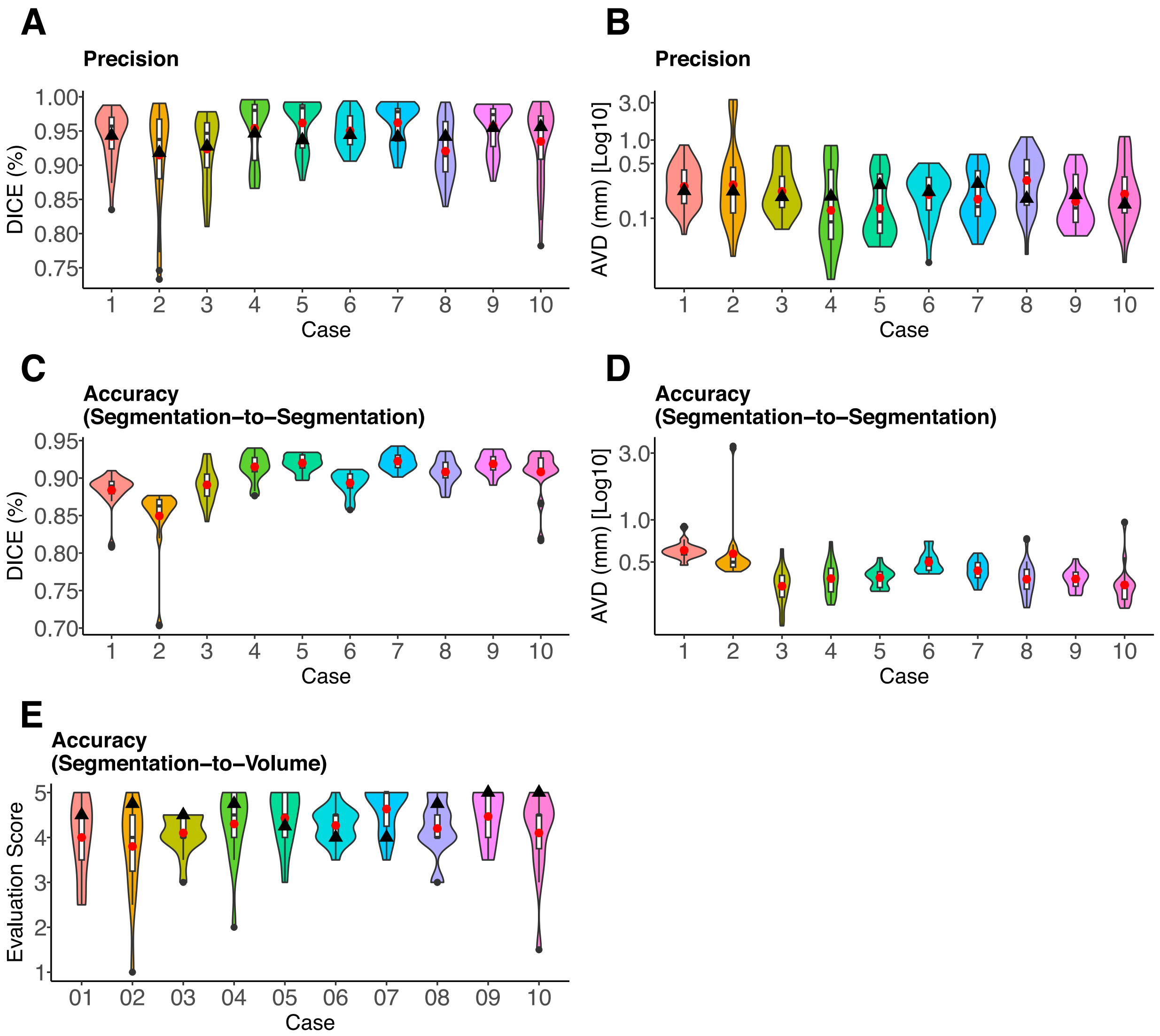

Fig. 6. Evaluation of precision (**a–b**) and accuracy (**c–e**) for each case, regardless of the method used. Segmentation precision was evaluated by (**a**) Dice coefficient (DICE) on the y-axis and (**b**) average Hausdorff distance (AVD) on the y-axis (Log10 scale); Segmentation-to-segmentation accuracy was evaluated by a segmentation-to-segmentation comparison using the segmentations of Wallner and colleagues as a reference with (**c**) DICE on the y-axis and (**d**) AVD on the y-axis (Log10 scale). (**e**) Overall accuracy determined by a segmentation-to-volume comparison, assessed by two independent and experienced radiologists who scored performance (1 = poor, 5 = perfect), as shown on the y-axis. (**a–e**) Cases 1–10 on the x-axis. The violin plots (colored) include a boxplot (white), with the mean value marked as a red point. Black points are outliers. The black triangle marks the mean value of the public segmentations done by Wallner and colleagues (Wallner et al., 2019).

Supplementary Fig. 1). In our cases, this was particularly true for regions toward the z-direction (i.e., cranio-caudal direction), such as the lower edge, alveolar crest, and cranial surface of the capitulum. This is consistent with the observation that with increasing voxel size in a CBCT, the deviation between the ground truth (laser scanner) and threshold-based segmentation increased in porcine mandibular cadavers (Dong et al., 2019). Furthermore, a *meta*-analysis showed that for alveolar bone height and thickness measurements, there is a direct correlation between errors and increases in voxel size (Y. Li et al., 2019). Finally, voxel size is considered the limiting factor for voxel-based modeling in CAS (F. Nysjö, Olsson, Malmberg, Carlbom, & Nyström, 2017). Therefore, one option would be to either perform CT image reconstruction aimed at an isotropic voxel size or, if this is not possible, to resample the voxel size to smaller and isotropic voxels (Noser et al., 2011). The latter would result in a loss of radiomic information (Shafiq-Ul-Hassan et al., 2017), but this would mitigate the potential error in bone segmentation. In other words, segmenting an additional voxel would not lead to an additional surface distance of 2 mm (as in Case 8) but only 0.75 mm (as in Case 10). Therefore, instead of the recommended slice thickness of < 1.25 mm in image reconstruction (van Baar et al., 2018), a slice thickness below < 1.0 mm should instead be targeted with isotropic voxels to improve inter-rater precision in segmentation. Aside from voxel size, reconstruction filters determine, in particular, whether the transitions in the HU between neighboring voxels are soft or sharp (Vergalasova, McKenna, Yue, & Reyhan, 2020). A soft reconstruction kernel leads to significant uncertainties due to the lack of a sharp bone–soft tissue margin (see Case 8, Fig. 5). This is consistent with a study that examined the accuracy of CT-based 3D bone surfaces and showed that sharp and bone reconstruction kernels yielded higher accuracy when generating 3D models through threshold-based segmentation than soft reconstruction kernels (Puggelli, Uccheddu, Volpe, Furferi, & Di Feo, 2019).

The clinical implications of these findings are as follows: First, CT image reconstruction should be optimized not for diagnostics but for CAS, taking subsequent steps into account, including segmentation with the generation of 3D surface models and actual surgical planning.

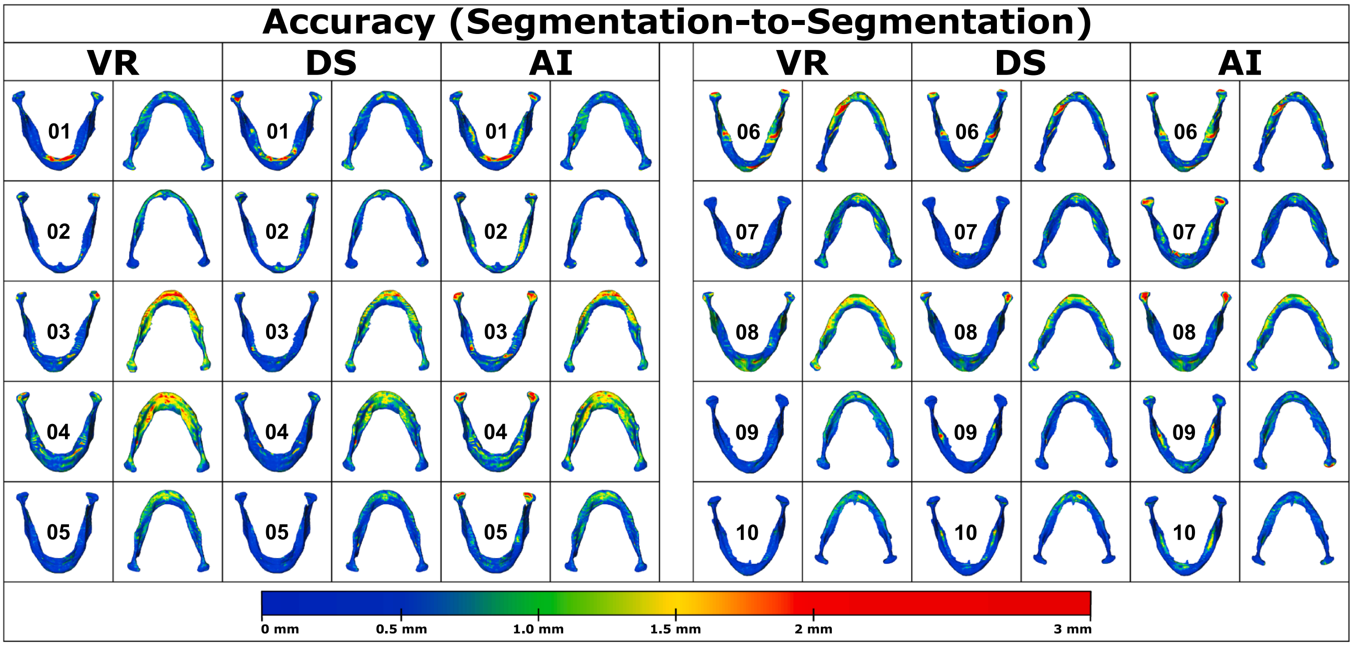

Fig. 7. Comparison of the accuracy (segmentation-to-segmentation) in virtual reality (VR), on a desktop screen (DS) and artificial intelligence (AI),whereby the public segmentations published by Wallner and colleagues (Wallner et al., 2019) (PUB_{DS}) served as a reference. The heatmaps were created by overlaying an average of the segmentations done by all five subjects and AI models over the reference per mandible, which show regions of high (red 2–3 mm), medium (green/yellow 0.5–2 mm), and low (blue 0–0.5 mm) variance.

Several considerations should be made, depending on the anatomical site involved. Patient-specific mandibular reconstruction plates (PSMRPs) have the advantage of reducing the rotational errors of the mandible compared to conventionally bent titanium reconstruction plates (Zeller et al., 2020). The external surface of the mandible in the segmented cases was very homogeneous (Figs. 5 and 7), which supports the idea that PSMRPs can be planned and positioned well on the outer surface. However, the deviation at the bottom of the lower edge of the mandible was much higher. Cutting guides are positioned over fitting surfaces that are complex or angled, have large surface areas or have lipping over the lower edge (Philippe, 2020). Possible positioning errors in the mandibular cutting guide may be explained by the fact that the lower edge shows the highest deviation. CAS applications where the foramen mandibulae is used as a landmark, such as in orthognathic surgery (Yang et al., 2011), should take into account that the segmentation shows high variability. Therefore, instead of the generated 3D surface model, the volume itself should be used as a landmark. Furthermore, volume and distance measurements should not be considered true values, especially for alveolar bone height (Y. Li et al., 2019). This is also true for the capitulum, which is difficult to segment due to its enclosure by the glenoid cavity (J. Li, Erdt, Janoos, Chang, & Egger, 2021; Wallner, Schwaiger, et al., 2019). The associated variability shows that CAS applications in the area of the temporomandibular joint (TMJ) (Memon, Wang, Hu, Egger, & Chen, 2020) should be particularly aware of its low precision.

Regardless, the usability of the method should not be ignored, as it could also influence accuracy and precision. In this context, we observed that the VR method created models a little larger than those created by DS, but not significantly (ANOVA, $p = 0.31$), while AI and the PUB_{DS} methods achieved the lowest volume (Fig. 3c). We assume that this was due to a blurring caused by the hardware limitation of the HMD that offered about 4.6 megapixels of resolution (combined resolution of 2880 × 1600 pixels), leading to difficulties in grasping the correct margin between bone and surrounding soft tissue. In fact, the HMD (HTC Vive Pro) used has a resolution of 13 pixels/degree, which is about six times lower than normal human vision (Cuervo, Chintalapudi, & Kotaru, 2018). The 27-inch Wide Quad High Definition (WQHD) monitor used for DS offered about 3.7 megapixels, with a sitting distance of 50 cm (eye/monitor distance). This resulted in a resolution of 41 pixels per degree (https://qasimk.io/screen-ppd/), which is over three times higher than the VR HMD used. As graphics power increases and display technology improves, the VR environment may become equivalent to the DS in terms of pixels per degree (Cuervo et al., 2018). Aside from visualization, the input device is important. Unlike a computer screen, where a mouse and keyboard are used, there have been concerns about whether controllers in VR would have the same accuracy as a mouse (Batmaz, Mutasim, & Stuerzlinger, 2020; Z. Li, Kiiveri, Rantala, & Raisamo, 2021). Precision grip controllers (three-finger configuration similar to holding a pen) have been shown to significantly reduce VR error rates (Batmaz et al., 2020) but have not been used for segmentation tasks. In contrast, power grip controllers (which encompass the entire hand) have been used successfully in segmentation (Ulbrich et al., 2023). Our study shows that precision grip controllers, such as the VR stylus, are well suited for use in segmentation tasks. This is consistent with a study showing that the combination of the VR stylus and controller is favored in medical marking tasks (Rantamaa et al., 2023).

Nevertheless, semi-automatic approaches have claimed to offer time advantages (Wallner, Schwaiger, et al., 2019) and have shown, high precision in threshold-based segmentation tasks involving a single investigator for different DS softwares (Lo Giudice et al., 2020). However, depending on the algorithms used, they led to unsegmented areas within the mandible or missing structures, showing a DICE of only 58.4–85.6 % compared to the gold standard of manual segmentation (Wallner, Schwaiger, et al., 2019). Furthermore, they are prone to errors in the midface area, such as the orbita with a complex and thin bone structure (Jansen et al., 2016). The results of the AI model included in this study are consistent with the reported concerns. Therefore, care must be taken to ensure that AI models perform well in the context of the local clinical setting. Considering this, VR could be a good alternative for saving time in manual segmentation. AI methods could be used for pre-segmentation, saving even more time, followed by correction of segmentation errors in VR. In the future, other steps in the CAS workflow could be performed in VR or combined with AI, making VR even more attractive for CAS workflows (Ulbrich et al., 2023). However, there are

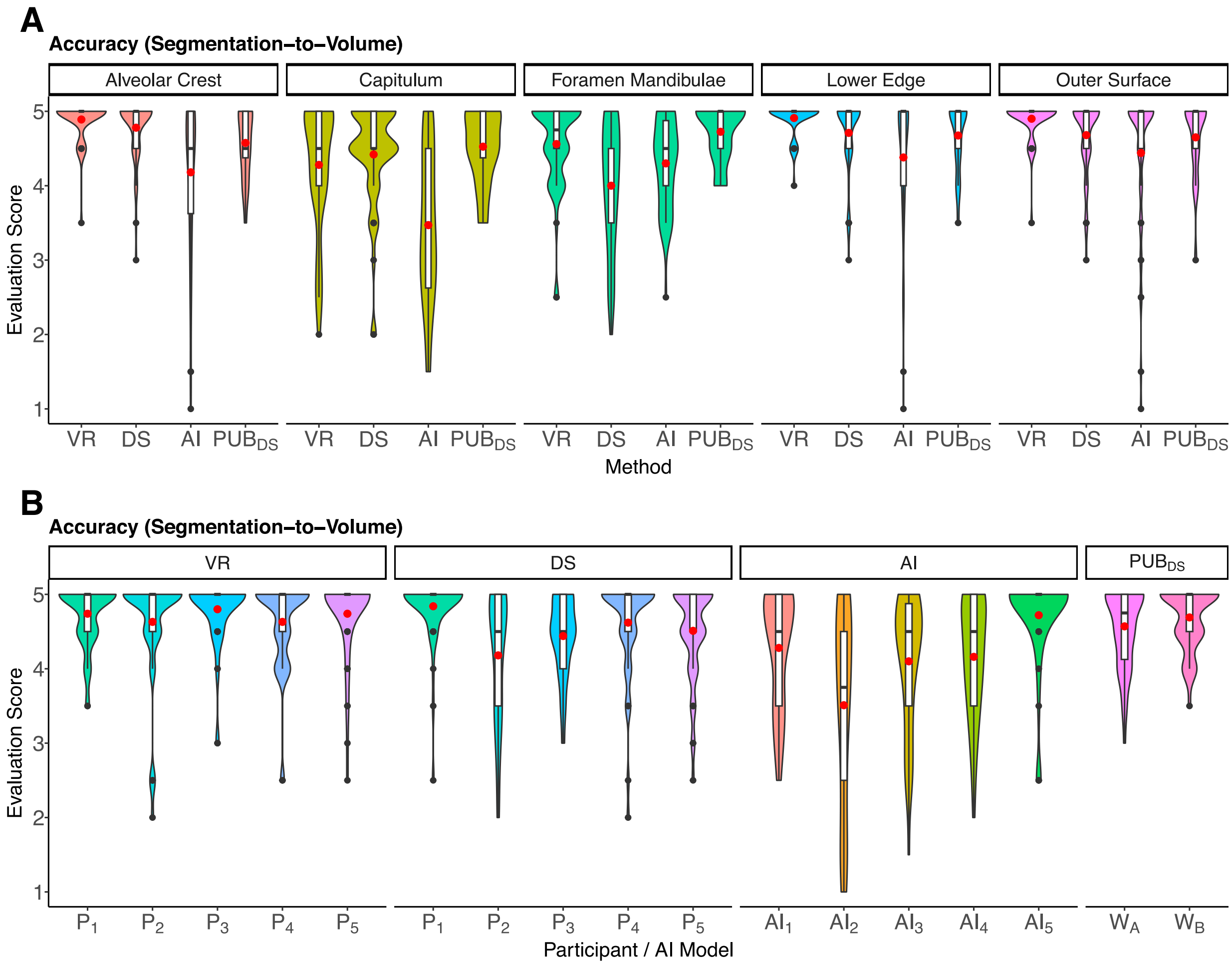

Fig. 8. Comparison of the overall accuracy (segmentation-to-volume) with an evaluation score on the y-axis, rated by two independent radiologists (1 = poor to 5 = perfect); (**a**): shows the rated accuracy regarding various anatomical sites of the mandible (alveolar crest, capitulum, foramen mandibulae, lower edge and outer surface) with the segmentation method used (VR, DS, AI, PUB$_{DS}$) on the x-axis; (**b**): shows the rated accuracy regarding the different segmentation methods with the results for each participant (P$_1$-P$_5$) in VR and DS, the different AI segmentation models AI$_1$-AI$_5$ and the two participants of the PUB$_{DS}$ method (W$_A$ and W$_B$) on the x-axis. The violin plots (colored) include a boxplot (white), with the mean value marked as a red point. Black points are outliers.

Table 3
The mean (standard deviation) results of the Likert questionnaire.

7-Point Likert Questions	VR (n = 50)	DS (n = 50)	Difference
I rate the segmentation by others as easily reproducible.	5.8 (0.8)	4.6 (1.1)	1.2 (1.3)
The working environment made segmentation easier for me.	6.2 (0.9)	4.4 (1.2)	1.8 (1.4)
I found this mandible sufficiently segmented for a CAS.	6.1 (1.0)	4.9 (1.3)	1.2 (1.5)
I found the time required for segmentation to be compatible with daily clinical practice.	6.0 (0.9)	4.4 (1.4)	1.6 (1.5)
I was able to grasp the anatomy of the mandible well in its entirety.	6.4 (0.9)	4.5 (1.4)	1.8 (1.5)
I was able to easily separate the temporomandibular joint from the skull base.	5.8 (1.1)	3.9 (1.5)	1.9 (1.7)
Filling cavities within the mandible was easy for me.	6.0 (1.0)	4.5 (1.3)	1.5 (1.7)
I was able to concentrate well during the segmentation process.	6.0 (0.9)	4.8 (1.1)	1.2 (1.2)
I was able to work precisely during the segmentation process.	6.0 (0.8)	4.7 (1.3)	1.3 (1.4)

two issues to consider before implementing VR in CAS workflows. First, the cost of VR systems, especially as certified medical software, should be considered. Second, the acceptance of VR by older professionals needs to be addressed.

The following limitations of the presented study should be considered. First, only edentulous mandibles were segmented. This meant that no artifacts occurred in the tooth area. However, an intra-oral scan of the teeth is always conducted and added to the segmented mandible as a surface model during the 3D planning of procedures involving teeth. Therefore, an augmented tooth model is used in most cases, which is why the use of an edentulous data set with edentulous mandible bones for segmentation evaluation makes sense. Furthermore, in the majority of the anatomical sites examined, artifacts played only a limited role. Second, a CT dataset in which not all images were fully optimized for CAS but were from a clinical routine was used (Wallner et al., 2019). However, this has the advantage of providing a good general estimate for clinical practice. Due to the public availability of these cases, future optimized applications could be assessed using them.

5. Conclusion

The clinical implications of our study are: depending on CT quality, method used, participants involved, and anatomical location, there is significant variability in the accuracy and precision of mandibular segmentation. In particular, the alveolar crest, capitulum, foramen mandibulae, and lower edge of the mandible showed remarkable variations in segmentation. This could negatively influence the subsequent steps of

CAS and lead to errors in applications and evaluation of patient-specific implants, cutting guides, robotics-guided scenarios, and augmented reality. Therefore, the claim to obtain a realistic 3D model of the mandible by CT and manual segmentation can only be achieved under certain conditions. A CT with a slice thickness of 1 mm or less should preferably be acquired and reconstructed with isotropic voxel size, while anisotropic voxels should be resampled. The observed variability should be considered when generating training data for AI and may explain, in addition to CT quality, why the AI exhibits human-like variability in the same anatomical regions. Although current AI models have perfect intramodel reliability, they have higher inter-model variability and are accompanied by invalid outliers making human review still necessary. In summary, the use of VR in manual segmentation showed high accuracy and precision overall while saving time, making it the preferred method over DS due to its good usability.

Declaration

Funding: Behrus Puladi was funded by the Medical Faculty of RWTH Aachen University as part of the Clinician Scientist Program. We acknowledge FWF enFaced 2.0 [KLI 1044, https://enfaced2.ikim.nrw/]; and KITE (Plattform für KI-Translation Essen) from the REACT-EU initiative [https://kite.ikim.nrw/, EFRE-0801977].

Institutional Review Board Statement: The study was approved by the Institutional Review Board (or Ethics Committee) of University Hospital RWTH Aachen (protocol code EK 471/20 and approval date 04.12.2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

CRediT authorship contribution statement

Lennart Johannes Gruber: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jan Egger:** Formal analysis, Writing – review & editing. **Andrea Bönsch:** Methodology, Writing – review & editing. **Joep Kraeima:** Formal analysis, Writing – review & editing. **Max Ulbrich:** Investigation, Writing – review & editing. **Vincent van den Bosch:** Validation, Writing – review & editing. **Ila Motmaen:** Investigation, Writing – review & editing. **Caroline Wilpert:** Validation, Writing – review & editing. **Mark Ooms:** Investigation, Writing – review & editing. **Peter Isfort:** Validation, Writing – review & editing. **Frank Hölzle:** Conceptualization, Resources, Writing – review & editing. **Behrus Puladi:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – review & editing, Visualization, Supervision, Project administration.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: The author Jan Egger is a member of the editorial board of the International Journal "Expert Systems with Applications".

Data availability

Data will be made available on request.

Acknowledgments

We thank Rainer Röhrig for scientific discussion regarding our study. We thank Justin Sutherland of Realize Med Inc. for providing a free license of the Elucis software. We thank Aidan Kehoe of Logitech for providing a free VR Ink Pilot Edition. We thank Michel Beyer, Xiaojun Chen, Antoine Coppens, Robert Ileșan, Tobias Pankert, Stefan Raith and Jiangchang Xu for applying their AI models and making the results available to us.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.eswa.2023.122275.

References

Batmaz, A. U., Mutasim, A. K., & Stuerzlinger, W. (2020). Precision vs. Power Grip: A Comparison of Pen Grip Styles for Selection in Virtual Reality. In *In 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)* (pp. 23–28). IEEE. https://doi.org/10.1109/VRW50115.2020.00012.

Bryan, F. W., Xu, Z., Asman, A. J., Allen, W. M., Reich, D. S., & Landman, B. A. (2014). Self-assessed performance improves statistical fusion of image labels. *Medical Physics, 41*(3), 31903. https://doi.org/10.1118/1.4864236

Christensen, A. M., Weimer, K., Beaudreau, C., Rensberger, M., & Johnson, B. The Digital Thread for Personalized Craniomaxillofacial Surgery, 23–45. https://doi.org/10.1007/978-1-4939-1532-3_2.

Cuervo, E., Chintalapudi, K., & Kotaru, M. (2018). Creating the Perfect Illusion. In M. Kim & A. Balasubramanian (Eds.), *Proceedings of the 19th International Workshop on Mobile Computing Systems & Applications* (pp. 7–12). New York, NY, USA: ACM. https://doi.org/10.1145/3177102.3177115.

Dong, T., Xia, L., Cai, C., Yuan, L., Ye, N., & Fang, B. (2019). Accuracy of in vitro mandibular volumetric measurements from CBCT of different voxel sizes with different segmentation threshold settings. *BMC Oral Health, 19*(1), 206. https://doi.org/10.1186/s12903-019-0891-5

Engelbrecht, W. P., Fourie, Z., Damstra, J., Gerrits, P. O., & Ren, Y. (2013). The influence of the segmentation process on 3D measurements from cone beam computed tomography-derived surface models. *Clinical Oral Investigations, 17*(8), 1919–1927. https://doi.org/10.1007/s00784-012-0881-3

Fourie, Z., Damstra, J., Schepers, R. H., Gerrits, P. O., & Ren, Y. (2012). Segmentation process significantly influences the accuracy of 3D surface models derived from cone beam computed tomography. *European Journal of Radiology, 81*(4), e524–e530. https://doi.org/10.1016/j.ejrad.2011.06.001

Gaêta-Araujo, H., Alzoubi, T., Vasconcelos, K.d. F., Orhan, K., Pauwels, R., Casselman, J. W., & Jacobs, R. (2020). Cone beam computed tomography in dentomaxillofacial radiology: A two-decade overview. *Dento Maxillo Facial Radiology, 49*(8), 20200145. https://doi.org/10.1259/dmfr.20200145

Gelasca, E. D., Ebrahimi, T., Farias, M., Carli, M., & Mitra, S. K. (2004). Towards Perceptually Driven Segmentation Evaluation Metrics, 52. https://doi.org/10.1109/CVPR.2004.465.

Gillot, M., Baquero, B., Le, C., Deleat-Besson, R., Bianchi, J., Ruellas, A., … Prieto, J. C. (2022). Automatic multi-anatomical skull structure segmentation of cone-beam computed tomography scans using 3D UNETR. *PLOS ONE, 17*(10), e0275033.

Greenberg, A. M. (Ed.) (2018). Digital Technologies in Craniomaxillofacial Surgery. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4939-1532-3.

Hofer, M., Strauß, G., Koulechov, K., & Dietz, A. (2005). Definition of accuracy and precision—evaluating CAS-systems. *International Congress Series, 1281*, 548–552. https://doi.org/10.1016/j.ics.2005.03.290

Ileșan, R. R., Beyer, M., Kunz, C., & Thieringer, F. M. (2023). *Comparison of Artificial Intelligence-Based Applications for Mandible Segmentation: From Established Platforms to In-House-Developed Software, 10*. https://doi.org/10.3390/bioengineering10050604

Jansen, J., Schreurs, R., Dubois, L., Maal, T. J. J., Gooris, P. J. J., & Becking, A. G. (2016). Orbital volume analysis: Validation of a semi-automatic software segmentation method. *International Journal of Computer Assisted Radiology and Surgery, 11*(1), 11–18. https://doi.org/10.1007/s11548-015-1254-6

Lalone, E. A., Willing, R. T., Shannon, H. L., King, G. J. W., & Johnson, J. A. (2015). Accuracy assessment of 3D bone reconstructions using CT: An intro comparison. *Medical Engineering & Physics, 37*(8), 729–738. https://doi.org/10.1016/j.medengphy.2015.04.010

Li, J., Erdt, M., Janoos, F., Chang, T., & Egger, J. (2021). Medical image segmentation in oral-maxillofacial surgery. In *Computer-Aided Oral and Maxillofacial Surgery* (pp. 1–27). Elsevier. https://doi.org/10.1016/B978-0-12-823299-6.00001-8.

Li, Y., Deng, S., Mei, L., Li, J., Qi, M., Su, S., … Zheng, W. (2019). Accuracy of alveolar bone height and thickness measurements in cone beam computed tomography: A systematic review and meta-analysis. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology, 128*(6), 667–679. https://doi.org/10.1016/j.oooo.2019.05.010

Li, Z., Kiiveri, M., Rantala, J., & Raisamo, R. (2021). Evaluation of haptic virtual reality user interfaces for medical marking on 3D models. *International Journal of Human-Computer Studies, 147*, Article 102561. https://doi.org/10.1016/j.ijhcs.2020.102561

Lo Giudice, A., Ronsivalle, V., Grippaudo, C., Lucchese, A., Muraglie, S., Lagravère, M. O., & Isola, G. (2020). One Step before 3D Printing-Evaluation of Imaging Software Accuracy for 3-Dimensional Analysis of the Mandible: A Comparative Study Using a Surface-to-Surface Matching Technique. Materials (Basel, Switzerland), 13(12). https://doi.org/10.3390/ma13122798.

Memon, A. R., Wang, E., Hu, J., Egger, J., & Chen, X. (2020). A review on computer-aided design and manufacturing of patient-specific maxillofacial implants. *Expert*

Review of Medical Devices, 17(4), 345–356. https://doi.org/10.1080/17434440.2020.1736040

Minnema, J., Ernst, A., van Eijnatten, M., Pauwels, R., Forouzanfar, T., Batenburg, K. J., & Wolff, J. (2022). A review on the application of deep learning for CT reconstruction, bone segmentation and surgical planning in oral and maxillofacial surgery. *Dento Maxillo Facial Radiology, 51*(7), 20210437. https://doi.org/10.1259/dmfr.20210437

Muehlematter, U. J., Daniore, P., & Vokinger, K. N. (2021). Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): A comparative analysis. *The Lancet. Digital Health, 3*(3), e195–e203. https://doi.org/10.1016/S2589-7500(20)30292-2

Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes, 15*(1), 210. https://doi.org/10.1186/s13104-022-06096-y

Noser, H., Heldstab, T., Schmutz, B., & Kamer, L. (2011). Typical accuracy and quality control of a process for creating CT-based virtual bone models. *Journal of Digital Imaging, 24*(3), 437–445. https://doi.org/10.1007/s10278-010-9287-4

Nysjö, F., Olsson, P., Malmberg, F., Carlbom, I., & Nyström, I. (2017). Using anti-aliased signed distance fields for generating surgical guides and plates from CT images. *Journal of WSCG, 25*, 11–20.

Nysjö, J. (2016). *Interactive 3D Image Analysis for Cranio-Maxillofacial Surgery Planning and Orthopedic Applications*.

Pankert, T., Lee, H., Peters, F., Hölzle, F., Modabber, A., & Raith, S. (2023). 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*(8), 1479–1488. https://doi.org/10.1007/s11548-022-02830-w

Pesapane, F., Volonté, C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: Ethical and regulatory issues in Europe and the United States. *Insights into Imaging, 9*(5), 745–753. https://doi.org/10.1007/s13244-018-0645-y

Philippe, B. (2020). Accuracy of position of cutting and drilling guide for sagittal split guided surgery: A proof of concept study. *The British Journal of Oral & Maxillofacial Surgery, 58*(8), 940–946. https://doi.org/10.1016/j.bjoms.2020.04.034

Powcharoen, W., Yang, W.-F., Yan Li, K., Zhu, W., & Su, Y.-X. (2019). Computer-Assisted versus Conventional Freehand Mandibular Reconstruction with Fibula Free Flap: A Systematic Review and Meta-Analysis. *Plastic and Reconstructive Surgery, 144*(6), 1417–1428. https://doi.org/10.1097/PRS.0000000000006261

Puggelli, L., Uccheddu, F., Volpe, Y., Furferi, R., & Di Feo, D. (2019). Accuracy Assessment of CT-Based 3D Bone Surface Reconstruction. In F. Cavas-Martínez, B. Eynard, F. J. Fernández Cañavate, D. G. Fernández-Pacheco, P. Morer, & V. Nigrelli (Eds.), *Lecture Notes in Mechanical Engineering. Advances on Mechanics, Design Engineering and Manufacturing II* (pp. 487–496). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-12346-8_47.

Qiu, B., van der Wel, H., Kraeima, J., Glas, H. H., Guo, J., Borra, R. J. H., … van Ooijen, P. M. A. (2021). Automatic Segmentation of Mandible from Conventional Methods to Deep Learning-A Review. *Journal of. Personalized Medicine, 11*(7). https://doi.org/10.3390/jpm11070629

Rantamaa, H.-R., Kangas, J., Kumar, S. K., Mehtonen, H., Järnstedt, J., & Raisamo, R. (2023). Comparison of a VR Stylus with a Controller, Hand Tracking, and a Mouse for Object Manipulation and Medical Marking Tasks in Virtual Reality. *Applied Sciences, 13*(4), 2251. https://doi.org/10.3390/app13042251

Shafiq-Ul-Hassan, M., Zhang, G. G., Latifi, K., Ullah, G., Hunt, D. C., Balagurunathan, Y., … Moros, E. G. (2017). Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. *Medical Physics, 44*(3), 1050–1062. https://doi.org/10.1002/mp.12123

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

Thambawita, V., Salehi, P., Sheshkal, S. A., Hicks, S. A., Hammer, H. L., Parasa, S., … Riegler, M. A. (2022). Singan-Seg: Synthetic training data generation for medical image segmentation. *PloS One, 17*(5), e0267976.

Torosdagli, N., Liberton, D. K., Verma, P., Sincan, M., Lee, J., Pattanaik, S., & Bagci, U. (2017). Robust and fully automated segmentation of mandible from CT scans. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)* (pp. 1209–1212). IEEE. https://doi.org/10.1109/ISBI.2017.7950734.

Ulbrich, M., van den Bosch, V., Bönsch, A., Gruber, L. J., Ooms, M., Melchior, C., … Puladi, B. (2023). Advantages of a Training Course for Surgical Planning in Virtual Reality for Oral and Maxillofacial Surgery: Crossover Study. *JMIR Serious Games, 11*, e40541.

Van Baar, G. J. C., Forouzanfar, T., Liberton, N. P. T. J., Winters, H. A. H., & Leusink, F. K. J. (2018). Accuracy of computer-assisted surgery in mandibular reconstruction: A systematic review. *Oral Oncology, 84*, 52–60. https://doi.org/10.1016/j.oraloncology.2018.07.004

Van Eijnatten, M., van Dijk, R., Dobbe, J., Streekstra, G., Koivisto, J., & Wolff, J. (2018). Ct image segmentation methods for bone used in medical additive manufacturing. *Medical Engineering & Physics, 51*, 6–16. https://doi.org/10.1016/j.medengphy.2017.10.008

Vergalasova, I., McKenna, M., Yue, N. J., & Reyhan, M. (2020). Impact of computed tomography (CT) reconstruction kernels on radiotherapy dose calculation. *Journal of Applied Clinical Medical Physics, 21*(9), 178–186. https://doi.org/10.1002/acm2.12994

Verhelst, P.-J., Smolders, A., Beznik, T., Meewis, J., Vandemeulebroucke, A., Shaheen, E., … Jacobs, R. (2021). Layered deep learning for automatic mandibular segmentation in cone-beam computed tomography. *Journal of Dentistry, 114*, Article 103786. https://doi.org/10.1016/j.jdent.2021.103786

Wallner, J., Hochegger, K., Chen, X., Mischak, I., Reinbacher, K., Pau, M., … Egger, J. (2018). Clinical evaluation of semi-automatic open-source algorithmic software segmentation of the mandibular bone: Practical feasibility and assessment of a new course of action. *PLOS ONE, 13*(5), e0196378.

Wallner, J., Mischak, I., & Egger, J. (2019). Computed tomography data collection of the complete human mandible and valid clinical ground truth models. *Scientific Data, 6*, Article 190003. https://doi.org/10.1038/sdata.2019.3

Wallner, J., Schwaiger, M., Hochegger, K., Gsaxner, C., Zemann, W., & Egger, J. (2019). A review on multiplatform evaluations of semi-automatic open-source based image segmentation for cranio-maxillofacial surgery. *Computer Methods and Programs in Biomedicine, 182*, Article 105102. https://doi.org/10.1016/j.cmpb.2019.105102

Wang, Z., Wang, E., & Zhu, Y. (2020). Image segmentation evaluation: A survey of methods. *Artificial Intelligence Review, 53*(8), 5637–5674. https://doi.org/10.1007/s10462-020-09830-9

Willemink, M. J., & Noël, P. B. (2019). The evolution of image reconstruction for CT-from filtered back projection to artificial intelligence. *European Radiology, 29*(5), 2185–2195. https://doi.org/10.1007/s00330-018-5810-7

Xu, J., Liu, J., Zhang, D., Zhou, Z., Zhang, C., & Chen, X. (2021). 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*, Article 104925. https://doi.org/10.1016/j.compbiomed.2021.104925

Yang, X., Hu, J., Zhu, S., Liang, X., Li, J., & Luo, E. (2011). Computer-assisted surgical planning and simulation for condylar reconstruction in patients with osteochondroma. *The British Journal of Oral & Maxillofacial Surgery, 49*(3), 203–208. https://doi.org/10.1016/j.bjoms.2010.03.004

Yu, S., Chen, M., Zhang, E., Wu, J., Yu, H., Yang, Z., … Lu, W. (2020). Robustness study of noisy annotation in deep learning based medical image segmentation. *Physics in Medicine and Biology, 65*(17), Article 175007. https://doi.org/10.1088/1361-6560/ab99e5

Zeller, A. N., Neuhaus, M. T., Weissbach, L. V. M., Rana, M., Dhawan, A., Eckstein, F. M., … Zimmerer, R. M. (2020). Patient-Specific Mandibular Reconstruction Plates Increase Accuracy and Long-Term Stability in Immediate Alloplastic Reconstruction of Segmental Mandibular Defects. *Journal of Maxillofacial and Oral Surgery, 19*(4), 609–615. https://doi.org/10.1007/s12663-019-01323-9

Zhao, F., & Xie, X. (2013). An overview of interactive medical image segmentation. *Annals of the BMVA, 2013*(7), 1–22.