

<https://doi.org/10.1038/s41746-025-01624-z>

# Multicentre evaluation of deep learning CT autosegmentation of the head and neck region for radiotherapy

Check for updates

Eric Pei Ping Pang<sup>1,2,15</sup>, Hong Qi Tan<sup>1,2,3,15</sup> ✉, Fuqiang Wang<sup>1,2</sup>, Jarkko Niemelä<sup>4</sup>, Gregory Bolard<sup>4</sup>, Susan Ramadan<sup>5</sup>, Timo Kiljunen<sup>5</sup>, Marta Capala<sup>6</sup>, Steven Petit<sup>6</sup>, Jan Seppälä<sup>7</sup>, Kristiina Vuolukka<sup>7</sup>, Ingrid Kiitam<sup>8</sup>, Danil Zolotuhhin<sup>8</sup>, Eduard Gershkevitch<sup>8</sup>, Kaisa Lehti<sup>9</sup>, Juha Nikkinen<sup>9,10</sup>, Jani Keyriläinen<sup>11,12,13</sup>, Miia Mokka<sup>11</sup> & Melvin Lee Kiang Chua<sup>1,2,14</sup>

This is a multi-institutional study to evaluate a head-and-neck CT auto-segmentation software across seven institutions globally. 11 lymph node levels and 7 organs-at-risk contours were evaluated in a two-phase study design. Time savings were measured in both phases, and the inter-observer variability across the seven institutions was quantified in phase two. Overall time savings were found to be 42% in phase one and 49% in phase two. Lymph node levels IA, IB, III, IVA, and IVB showed no significant time savings, with some centers reporting longer editing times than manual delineation. All the edited ROIs showed reduced inter-observer variability compared to manual segmentation. Our study shows that auto-segmentation plays a crucial role in harmonizing contouring practices globally. However, the clinical benefits of auto-segmentation software vary significantly across ROIs and between clinics. To maximize its potential, institution-specific commissioning is required to optimize the clinical benefits.

The transformative roles and applications of artificial intelligence (AI) in healthcare has led to rapid development of the technology and exploration of its potentials in advancing care in patients<sup>1,2</sup>. In radiotherapy, AI is expected to play an important role in almost all the steps of a radiotherapy workflow from CT simulation to treatment delivery<sup>3</sup>. Prior to treatment planning, clinicians have traditionally manually contoured organs at risk (OARs) and target volumes on CT scans. This time-consuming, yet indispensable, procedure ensures the accurate assessment of radiation dose to normal tissues and target volume<sup>4</sup>. From an atlas-based contouring to machine learning and now towards deep learning (DL)-based contouring<sup>5,6</sup>, there is an increased adoption of auto-segmentation in the clinic due to the accurate and efficient output and numerous clinically ready commercial DL-based auto-segmentation models<sup>7,8</sup>. This was demonstrated in a recent

study by E. Gibbons et al., which found that DL auto-segmentation has outperformed the atlas-based method in the contouring accuracy of OARs in head-and-neck, thoracic and pelvic regions<sup>9</sup>.

Specifically for head-and-neck cases, several studies<sup>10–12</sup> have highlighted the resource intensive process of manual delineation due to the complexity and extensive number of OARs and lymph node (LN) levels involved in this anatomical region. In addition, the heterogenous and inconsistent contouring among expert physicians has also been widely reported<sup>13–15</sup>. Hence, the industry interest towards the development and use of AI-based annotations of medical images has increased significantly over the last decade to solve the logistics and inter-observer variability challenges. Several studies have reported on the numerous clinical benefits of the auto-segmentation algorithms. X. Ye et al. and J. J. Lucido et al. have reported

<sup>1</sup>Division of Radiation Oncology, National Cancer Centre Singapore, Singapore, 158683, Singapore. <sup>2</sup>Oncology Academic Clinical Programme, Duke-NUS Graduate Medical School, Singapore, 169857, Singapore. <sup>3</sup>Division of Physics and Applied Physics, School of Physical and Mathematical Science, Nanyang Technological University, Singapore, Singapore. <sup>4</sup>MVision Ai, c/o Terkko Health hub, FI-00290 Helsinki, Finland. <sup>5</sup>Docrates Cancer Center, 00180 Helsinki, Finland. <sup>6</sup>Department of Radiotherapy, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, Netherlands. <sup>7</sup>Kuopio University Hospital, Center of Oncology, FI-70210 Kuopio, Finland. <sup>8</sup>North Estonia Medical Centre, 13419 Tallinn, Estonia. <sup>9</sup>Oulu University Hospital, Department of Oncology and Radiotherapy, 90220 Oulu, Finland. <sup>10</sup>Research Unit of Health Sciences and Technology, Faculty of Medicine, University of Oulu, 90220 Oulu, Finland. <sup>11</sup>Turku University Hospital, Department of Oncology and Radiotherapy, FI-20521 Turku, Finland. <sup>12</sup>Turku University Hospital, Department of Medical Physics, FI-20521 Turku, Finland. <sup>13</sup>University of Helsinki, Faculty of Science, Department of Physics University of Helsinki, FI-00014 Helsinki, Finland. <sup>14</sup>Division of Medical Sciences, National Cancer Centre Singapore, Singapore, 169610, Singapore. <sup>15</sup>These authors contributed equally: Eric Pei Ping Pang, Hong Qi Tan.

✉ e-mail: [tan.hong.qi@nccs.com.sg](mailto:tan.hong.qi@nccs.com.sg)

significant time saving benefits in the auto-segmentation of OARs in head-and-neck cancer where more than 76%-time reduction has been observed<sup>18,16</sup>. L. J. Stapleford et al. demonstrated that atlas-based LN segmentation reduced inter-observer contouring variation<sup>17</sup> and similar findings have also been observed in other treatment sites such as in breast cancer<sup>18</sup>.

There are several studies regarding DL-based auto-segmentation in head-and-neck cancers but most of the works are limited to algorithm development, performance quantification and are also a largely single-institutional study<sup>19–22</sup>. At the time of writing, only one study has reported a non-randomised multi-centre study using a stratified OAR segmentation system (SOARS) to automatically delineate 42 head-and-neck OARs<sup>8</sup>. They evaluated SOARS in five external institutions and found up to 90% clinical workload reduction and a smaller inter-observer variation in contouring. Despite the reported clinical benefits, it is pertinent to acknowledge potential observers' fatigue, which may be experienced by the limited human experts in the study due to the need to review the OAR contours from many patients and not all the 42 region-of-interests (ROIs) were evaluated for all five institutions due to varying institutional practice in OAR contouring. These are potential confounders in the study. Lastly, all the five institutions are located within a single country and therefore the findings may not be representative of the clinical benefits in a global setting. Our group had previously investigated the accuracy and efficiency of DL-based CT auto-segmentation tools for prostate cancer cases and have reported improved consistency and time saving benefits of the tool<sup>23</sup>. The present HARMONY (HeAd neck Rapid deep-learning auto-segmentation tool – a Multi-clinic evaluation study) study adopts a multi-institutional approach with randomization of scans to investigate the clinical benefits of a commercial auto-segmentation algorithm in head-and-neck CT scans.

The proposed study involves seven institutions located across different continents and is structured into two phases. The first phase assesses the time saving and the auto-segmentation accuracy of the auto-segmentation algorithm within each individual institution using their own internal CT scans. The second phase investigates the time saving and the reduction in inter-observer contouring variability among the seven institutions by pooling scans from different institutions. Through the proposed study design and the inclusion of institutions from distinct geographical locations, we aimed to quantify a more representative understanding of the clinical benefits of auto-segmentation software in the global setting, and to identify heterogeneity in these benefits across institutions. Additionally, this study also includes evaluates the clinical benefits of auto-segmenting the various LN levels, an aspect, which has not been reported previously in a multi-center setting.

## Results

### Data characteristics and study design

The schematic of this study is illustrated in Fig. 1. The aim of phase one was to assess the auto-segmentation qualities of all the 18 ROIs in the 35 CT scans (5 CT scans per clinic × 7 clinics). The assessment was quantified via two approaches. The first objective was to determine the absolute value of the contour similarity metrics between auto-segmentation and manual contours. The second objective focused on comparing the differences in similarity metrics between auto-segmentation versus manual contours and edited versus manual contours. If no statistically significant differences were detected between the metrics, it would imply that the auto-segmentation closely approximates clinically acceptable standards, requiring minimal editing. The aim of phase two was to assess the reduction in inter-observer contouring variability between the manual and the edited contours (when users were presented with the auto-segmentations). This was evaluated among the seven clinics on seven unique CT scans for all the 18 ROIs. The time saving clinical benefits were evaluated in both phases of the study. This analysis aimed to investigate whether the time-saving benefits were observed when the clinicians contoured internal CT scans from their own institution compared to external CT scans from other institutions.

### Contouring time saving in phase one and two

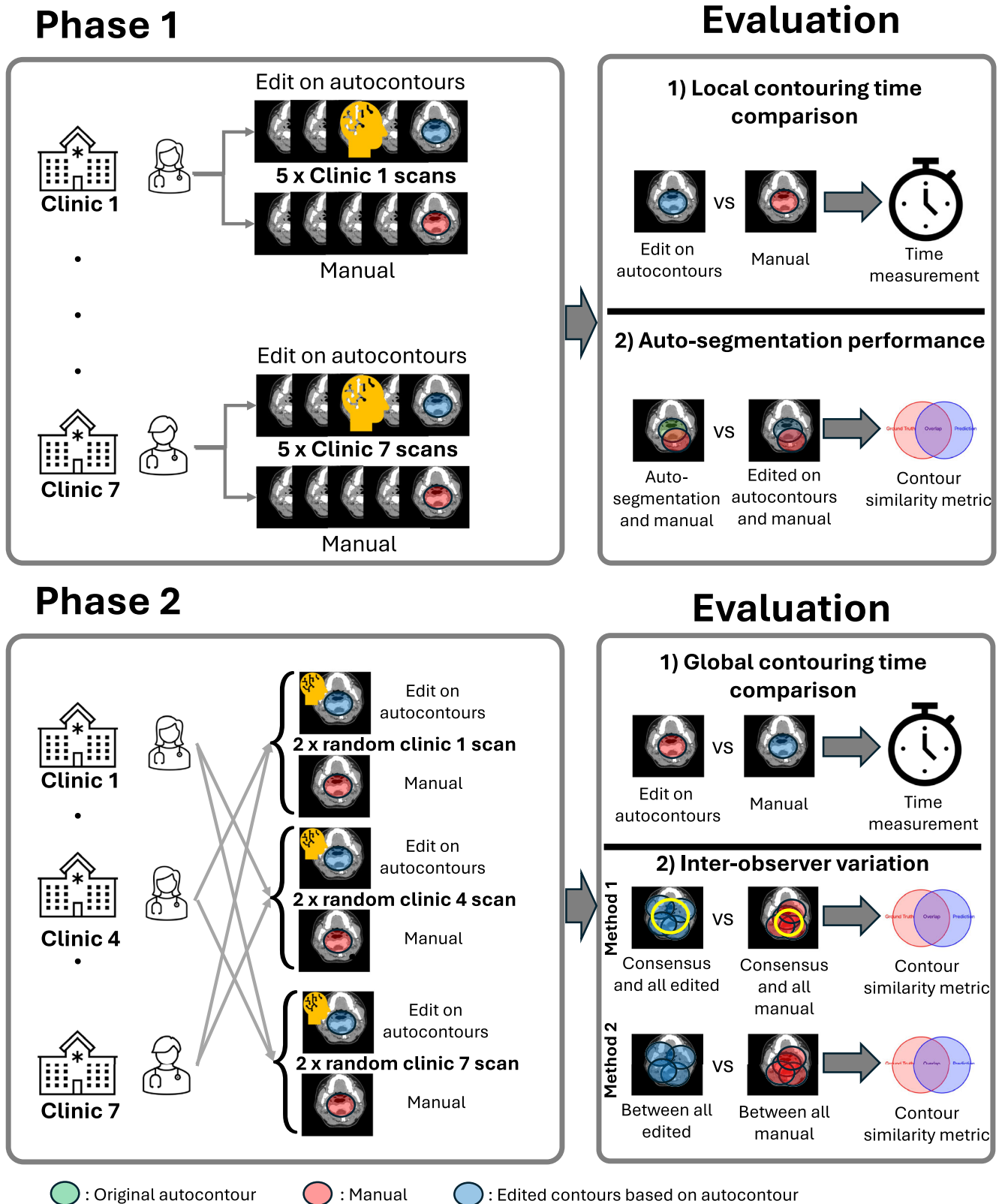
The contouring time saving results for phases one and two are shown in the left and right panels of Fig. 2. Figure 2a, b show the box plots of the manual contouring and editing time for all the ROIs in phase one and two, respectively. Using the Wilcoxon signed-rank test, statistically significant time reduction was observed for majority of the ROIs in both phases, except for LN IA, IB\_R, IB\_L, III\_R, IVA\_L, IV\_A\_R, IVB\_L and IVB\_R (indicated by \*\* in the figures). The percentage time reductions after the use of auto-segmentation software are shown in Fig. 2c, d. A positive value meant a shorter time for editing compared to manually contouring the ROI and therefore, it indicated that the ROI benefitted from the use of an auto-segmentation software. In terms of average time savings, the parotid glands and brachial plexus benefitted the most with more than 50% reduction in contouring time. Conversely, the time required to edit LN IVB contours took longer than manually contouring them from scratch. This observation was consistent in both phase one and two. Lastly, the total time reduction in contouring for all the ROIs in each individual clinic is shown in Fig. 2e, f. The time saving was non-uniform across the clinics. The “lion” clinic benefitted the most in terms of absolute time saving while the “duck” clinic benefits the least. The total segmentation time in phase one decreased from 90.7 ± 52.2 min to 52.2 ± 22.7 min, and in phase two decreased from 74.4 ± 56.2 min to 37.8 ± 28.0 min. Overall, the segmentation time decreased across both phases from 81.9 ± 54.7 min to 44.4 ± 26.5 min, which amounts to about 46% time saving.

### Auto-segmentation performance in phase one

Figure 3a, b shows the box plots of Dice score and HD of auto-segmentation versus manual contour (blue) and edited contour versus manual contour (orange). The corresponding relative difference between the Dice and HD is shown in Fig. 3c, d. Using the Wilcoxon signed-rank tests, all the improvements in the Dice score after editing on the auto-segmentation were found to be statistically significant. The improvements in HD were less conclusive with the parotid (left and right), brachial plexus (left and right) and LN IVB (left and right) showing no statistically significant improvement in contouring similarity after editing on the auto-segmentation. It was interesting to note that the Dice score between the edited and manual contours fall below 0.80 for most ROI with the brachial plexus having the lowest average Dice score of below 0.60.

### Inter-observer variation in phase two

The inter-observer contouring variation among the manual and edited contours is shown in Fig. 4. Figure 4a, c shows the Dice score and HD for the comparison between the manual contours and its consensus (blue) and the edited contours and its consensus (orange). Using the Mann–Whitney U tests, all the Dice score (HD) for the ROIs show statistically significant increase (decrease) in the edited contours. This shows that the inter-observer variation in contouring decreases when clinician edited the auto-segmentation results, when compared to manually contouring from scratch. The results using the pairwise comparison approach is shown in Supplementary Table 1 and similar reduction in inter-observer variation is observed in all ROIs as well. The left and right brachial plexus show the greatest decrease in inter-observer variations in the edited contours. The manual and edited contours for brachial plexus, LN IVA, IVB and parotid ROIs from all the clinicians are compiled and displayed in Fig. 5. Additional visual comparisons of all the ROIs are shown in Supplementary Fig. 1–6. The edited contours are clearly a lot more consistent among the clinicians as shown in the right panel of the figure. The great improvement in consistency of brachial plexus contouring is also evident in the figure. The results are also stratified according to the clinics and the similarity metrics of all the ROIs are shown in Fig. 4b, d. The “lion” clinic benefitted the most in term of increased agreement to the consensus contour after editing on the auto-segmentation.



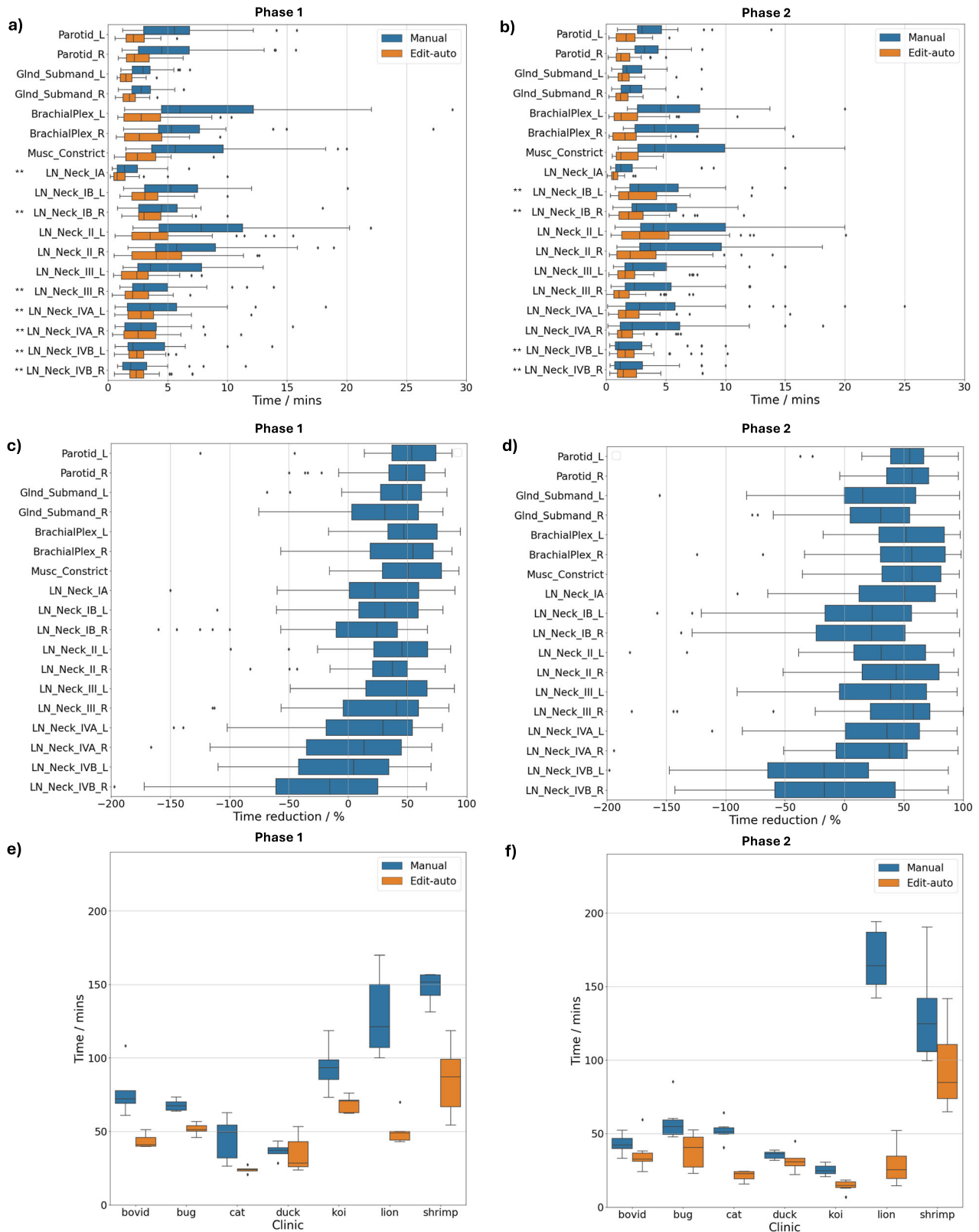
**Fig. 1 | Schematics of the study.** This figure shows the design of the two phases study. The first phase compares the manual and AI-generated contours within each institution while the second phase compares across different institutions. The

pooling and exchange of scans in phase two is to reduce institutional CT scan-specific bias when evaluating the time reduction and inter-observer contouring variation. AI artificial intelligence, CT computed tomography.

**Heterogeneity in performance across clinic and ROI**

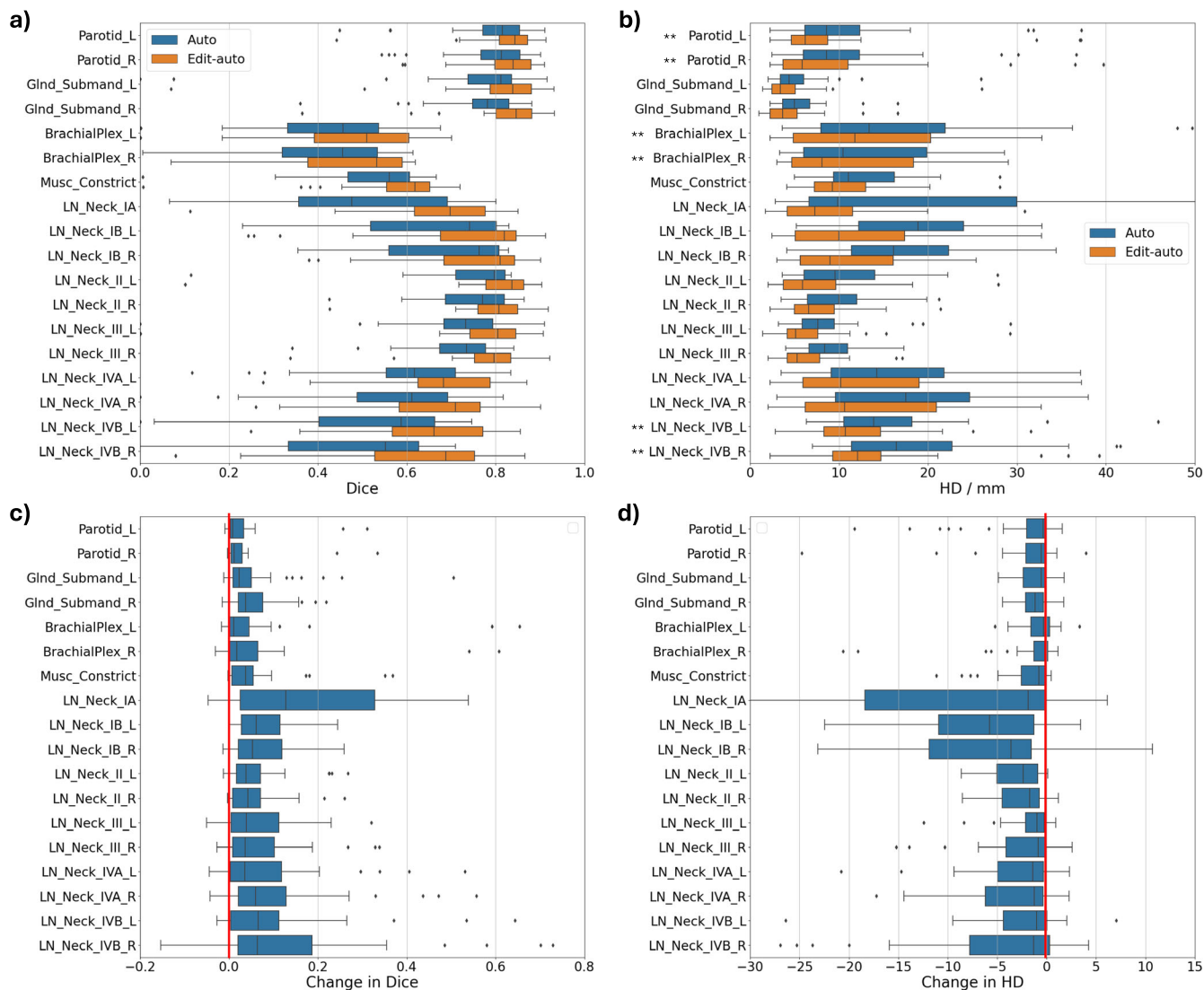
Figure 6 shows the details of the time saving and contour similarity from Figs. 2 to 4 as a function of the clinic and the ROI. The color bar represents the percentage reduction in time or improvement in similarity metrics,

with red representing the benefits of edited contours and blue indicating a lack of improvement. The size of the marker is correlated to the contouring time and similarity metrics evaluated on the manual contours. The heterogeneity in the results across the clinics and ROIs is evident in



**Fig. 2 | Comparison of the contouring time for phase one and two of the studies.** **a, b** The time taken to contour each ROI for phase one and two, respectively. **c, d** The time reduction for each ROI between manual and edit-auto. **e, f** The total time taken to finish contouring all the 18 ROIs for the CT scans from each clinic for phase one and two. Each boxplot in **(b)** represents the results from the five CT scans contoured by a single clinician in the clinic, while each boxplot in figure **d** shows the contouring

time of six different clinicians on a single CT scan from a different clinic. The blue and orange boxplots represent the time taken to manually contour the ROI from scratch and the editing time of an AI-generated contour. (\*\*) in **(a)** and **(b)** denotes no statistically significant difference in the contouring time between the manual and edit-auto ROI. ROI region-of-interest, CT computed tomography, AI artificial intelligence, LN lymph node, L left, R right.



**Fig. 3 | Comparison of contour similarity of the manual contour versus auto-segmentation and manual contour versus edited ROI in phase one. a, b** The Dice and HD of the comparison respectively. The manual versus auto-segmentation and manual versus edited ROI results are shown as blue and orange box plots respectively. The difference between the Dice of edited ROI and auto-

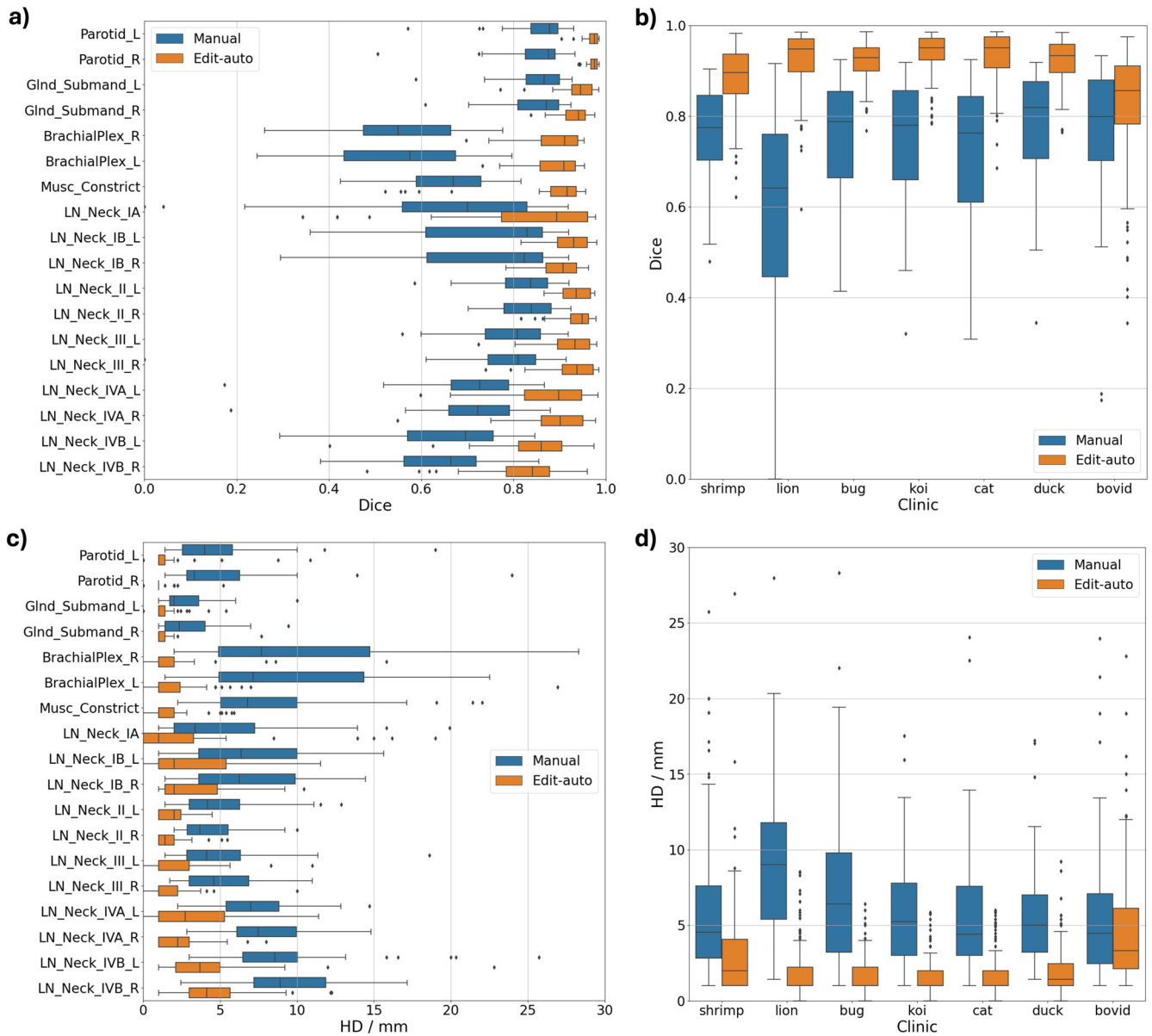
segmentation versus manual is shown in (c). Similarly, the difference between HD of edited ROI and auto-segmentation versus manual is shown in (d). (\*\*) in (b) denotes no statistically significant difference between the HD of manual versus auto-segmentation and manual versus edit-auto ROI. The red lines in (c) and (d) represent zero editing.

this figure. The “lion” clinic benefitted the most in terms of contouring time (in both phase one and two) and agreement to the consensus contours as indicated by the intense red markers. The edited contours from the “bovid” clinic benefitted the least in terms of inter-observer variation with LN II, LN III and the submandibular glands showing a poorer agreement to the consensus contour. Both the “bovid” and “duck” clinics benefitted the least in terms of time saving with multiple ROIs in both phases showing an increased editing time compared to manual contouring time (shown as blue markers). In both phases, all clinics except the “lion” take a longer time to edit than to manually contour the LN IVA and IVB ROIs (indicated by blue markers).

### Discussion

In this study, we have shown for the first time the heterogeneity in clinical benefits of using a head-and-neck auto-segmentation software in a multi-institutional and inter-continental setting. The time saving, inter-observer contouring variation and auto-segmentation accuracy with respect to manual contours in each local institution are quantified through a two phases study design. The results in Fig. 2e, f show that all

clinics benefitted from a contouring time reduction using the auto-segmentation software. This observation is consistent across both phases regardless of whether the clinicians are contouring on scans within or outside their institution. The total time saving in both phases is 46%, which is in close agreement to other single institutional auto-segmentation studies<sup>24–26</sup>. However, a closer analysis of the time saving stratified by ROIs and clinics in Figs. 2, 6a, b reveals a very heterogeneous performance, which sheds light on the actual clinical benefits in a global setting. As depicted in Fig. 2a, b, the parotid glands and brachial plexus experienced the most significant time savings, exceeding 50%. However, not all ROIs benefitted to the same extent, since LN IA, IB, III, IVA and IVB showed no statistically significant differences in contouring time between manual contouring and editing the auto-segmentation. Figure 6a, b elucidate the reason behind this finding, revealing that the time savings for these ROIs vary significantly between the institutions, as evidenced by the mixture of blue and red markers in the horizontal rows. The results show that in the majority of the clinics, manually contouring these ROIs is faster than editing the auto-segmentations. This is especially true for LN IVA and IVB in both



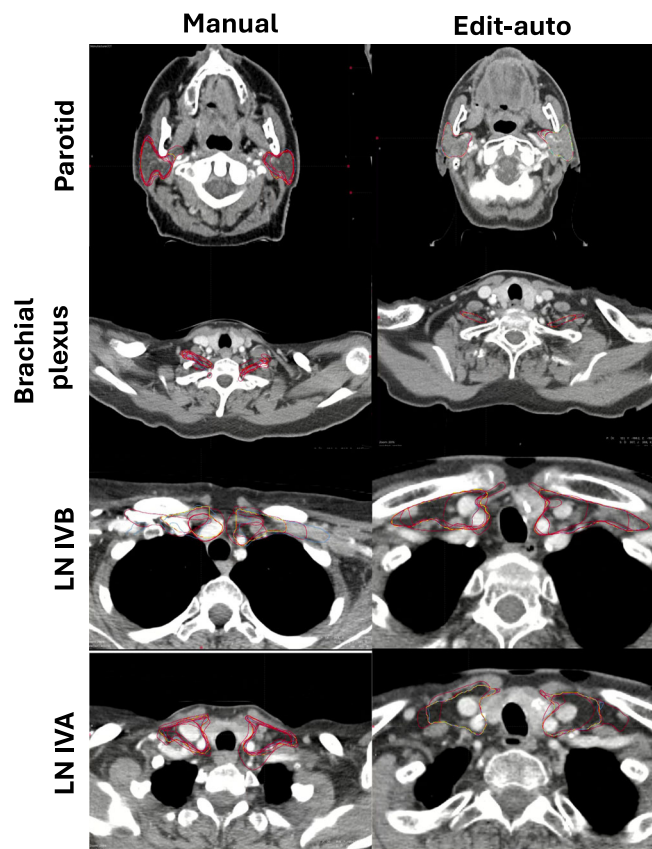
**Fig. 4 | Comparison of the contour similarity of manual versus consensus and edited ROI versus the consensus contours. a, c** The Dice and HD results between the manual and edited ROI and the consensus contour, respectively, for each ROI. **b, d** Similar results for all the ROIs in each of the seven unique CT scans. The blue

and orange boxplots represent similarity metrics of manual versus consensus and edited ROI versus consensus contours respectively. HD 95<sup>th</sup> percentile Hausdorff distance, ROI region-of-interest, CT computed tomography, LN lymph node, L left, R right.

phases. The only distinct exception is the “lion” clinic, which shows time saving in all the ROIs with the use of auto-segmentations in both phases (except for LN IA in phase one). In fact, time saving of more than 75% can be observed in multiple ROIs such as brachial plexus, constrictor muscle and the various LN levels in the “lion” clinic. In all, the differential time saving benefits across the ROIs and clinics can be attributed to the intricate interplay between the structure complexity, experience of the clinician and the contouring software in the institution. Interestingly, these results also show that the use of auto-segmentation in a clinic can be optimized by omitting the auto-segmentation of certain ROIs, which prove to take a longer time to edit than to manually contour. Ideally, this optimization process should take place during commissioning of the software and a proper assessment needs to be conducted since it is impossible to predict a priori, which ROI will benefit from it.

The reduction in the inter-observer contouring variability in the edited contours is evident in the results of the phase two study as shown

in Fig. 4. All the ROIs in Fig. 4a, c show a statistically significant improvement in the Dice and HD for the edited contours and its consensus compared to the manual counterpart. The brachial plexus shows the most significant reduction in inter-observer variability, as illustrated in Fig. 5, where the edited contours have greater consistency compared to the manual contours. This consistency is crucial for accurate OAR dose calculations, especially in the treatment of lower neck nodes and gross tumour volumes (GTV). Additionally, all the LN contours show an enhanced consistency, with LN IVB showing the highest average improvement. We also evaluated the inter-observer variation in combined LN groups, which are used in clinical practice. The results in three LN groups (I + II + III, II + III, II + III + IV) are shown in Supplementary Table 2 and 3 where an average Dice score of close to 0.90 in the pairwise comparison. Even though Ye et al. has shown a reduced inter-observer contouring variability between multiple institutions in head-and-neck OARs<sup>8</sup>, this is the first study to report similar findings in the contouring of the LN clinical target volumes.



**Fig. 5 | Visual comparison of the manual and edited contours in phase two.** This figure shows a comparison of the six manual and edited contours on each CT scans. The red contours represent the manual and edited delineations made by each clinician, while the blue contours represent the auto-segmentation. The yellow contours represent the consensus contour. In the right panel, the blue and yellow contours are less visible due to substantial overlap with the red edited contours. CT computed tomography, LN lymph node.

Figure 4b, d show that “lion” clinic benefitted the most from the improved contouring consistency. A closer look at Fig. 6c, d show that brachial plexus, LN IA and IB improved the most in this clinic. Visually, apart from the “lion” clinic, the rest of the clinics has reasonably similar improvement in the contouring consistency. The vast improvement for “lion” compared to the rest of the clinics could be due to a difference in training, experience and cancer type prevalence within the clinic and country. Auto-segmentation, therefore, could play an important role in harmonizing ROI contouring between different institutions and lower the learning curve for institution with limited experience.

The auto-segmentation accuracy with respect to the manual contours in each institution is shown by the blue box plot in Fig. 3a. Following the editing process, all the ROIs show a statistically significant increase in the Dice score when compared to the manual contours. This indicates that editing is still required for all contours to achieve clinically acceptable quality in each clinic. However, using a Dice threshold of 0.80 (from TG132<sup>27</sup>) for an acceptable contour agreement, half of the ROIs do not agree with the manual contours even after editing (orange box plot). The only contours, which have an average Dice score exceeding 0.80 after editing are the parotid glands, submandibular glands and LNs IB, II and III. This implies that the institutional practices for contouring the other ROIs generally diverge from the DAHANCA guidelines used in the training dataset, or it may indicate that the auto-segmentation process did not perform optimally. These results highlight the actual clinical situation when a purchased auto-segmentation software is often trained on data from other centers.

However, interestingly, the results show that the edited contours are eventually closer to the auto-segmentation compared to the original manual contours. This again highlights the value of auto-segmentation in harmonizing the contouring practice globally, which plays an important role when performing data pooling analysis in a multi-institutional clinical trial<sup>28</sup>.

There are two limitations in our study. Firstly, several ROIs were omitted from this study, including the optic apparatus, retro-pharyngeal constrictors, brainstem, spinal cord, esophagus, larynx and oral cavity. The consensus guideline recommends contouring at least 40 OARs in head-and-neck cancer patients<sup>29</sup>, and it would certainly improve the clinical relevance of the study if more OARs were evaluated. However, the resources required from each clinic (contouring, timing and the institutional improvement approval), are important considerations. Consequently, this study primarily focused on the LN level contouring due to a lack of multi-institutional auto-segmentation research in this specific area. Secondly, the auto-segmentation performance in the presence of dental filling or obturator artifacts was not assessed in this study and the results in this study only applied to pre-operative head-and-neck CT scans as the post-operative CT scans were excluded.

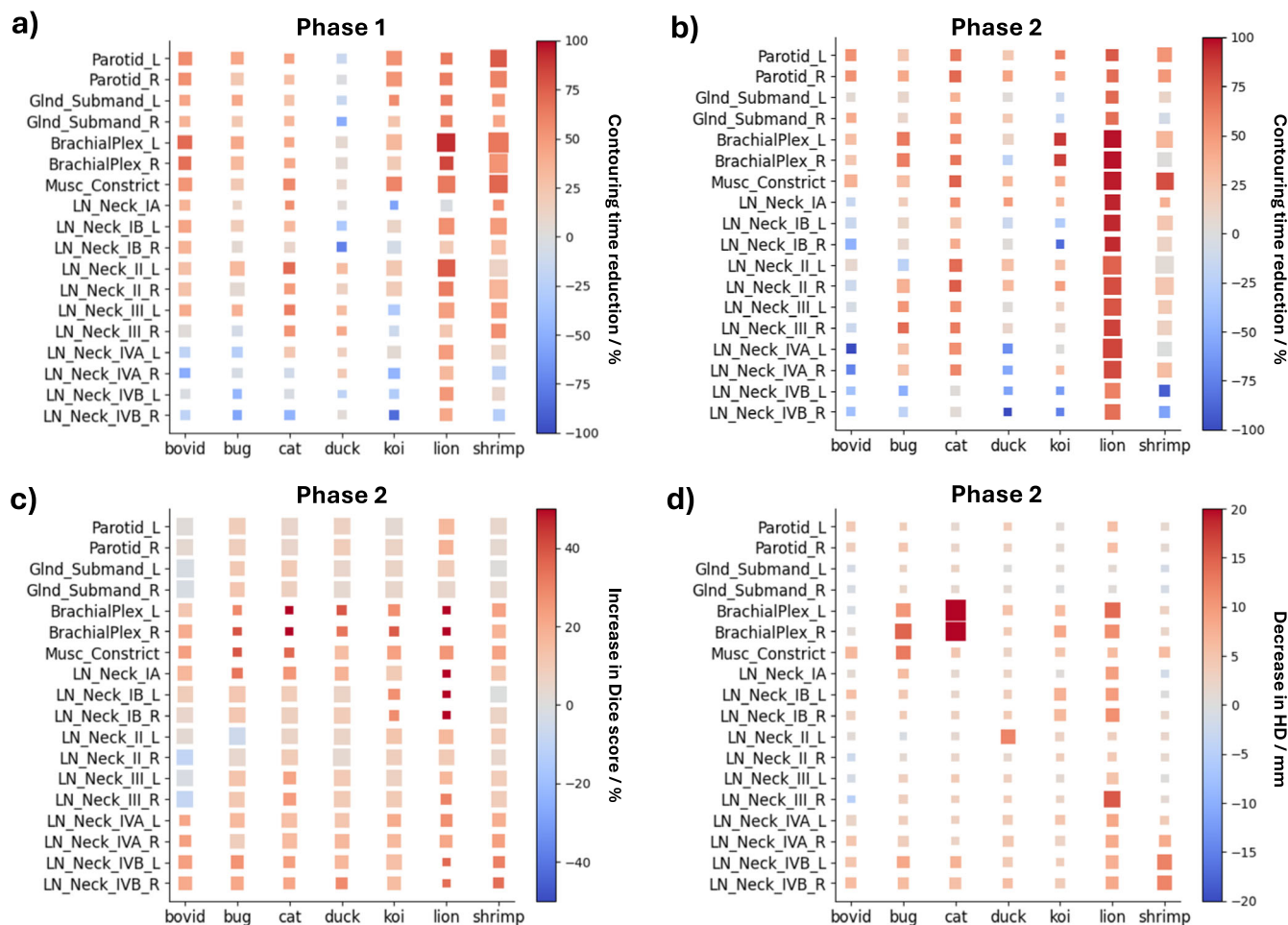
## Methods

### Data characteristics

This was a two-phase multi-centers retrospective randomized study involving seven institutions to evaluate the accuracy, efficiency and clinical benefit of a DL-based auto-segmentation in comparison to the conventional manual contouring processes. This study was conducted retrospectively and did not involve human subjects. A total of seven clinics participated in this study: 1) National Cancer Centre Singapore, Singapore 2) Docrates Cancer Center, Finland, 3) Erasmus MC Cancer Institute, Rotterdam, Netherlands, 4) Kuopio University Hospital, Finland, 5) North Estonia Medical Centre, Estonia, 6) Oulu University Hospital, Finland, 7) Turku University Hospital, Finland. The identities of the clinics were anonymized in the results of this study. They were labelled as bovid, bug, cat, duck, koi, lion and shrimp. Ethics clearance was obtained at each of the institutions based on the regulatory guidelines at the respective participating clinical site. Informed consent was waived as this study used only anonymized CT scans without associated clinical information. Inclusion criteria for the study consisted of pre-operative head-and-neck cancer patients who underwent CT simulation scans, with no restriction on tumour (T) or nodal (N) classification. Conversely, post-operative cases were excluded from the study.

### Auto-segmentation software and algorithm

The segmentation model was developed using a 3D encoder-decoder U-Net architecture, similar to the approach described in our previous CT segmentation work<sup>30</sup>. The model was trained on a dataset of approximately 500 CT scans from head-and-neck cancer patients, with manual segmentations performed by experienced radiation oncologists. OARs were delineated following multiple guideline sources, including DAHANCA, Brouwer et al., and Scoccianti et al.<sup>29,31,32</sup>, while LN levels were contoured according to the 2013 update of the international consensus guidelines by DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, and TROG<sup>33</sup>. Data augmentation techniques including random rotations, translations, and intensity transformations were applied to improve model generalization. The encoder utilized a ResNet-type backbone pre-trained on ImageNet (Stanford Vision Lab, Stanford University, Stanford, CA, USA), while the decoder comprised multiple DenseNet blocks. The model was optimized using a combination of Dice loss and weighted cross-entropy loss. Training was performed for roughly 300 epochs using the Adam optimizer and a ReduceLROnPlateau scheduler. All preprocessing and training was implemented in Python (Python Software Foundation, Wilmington, DE, USA) using the PyTorch framework. The final model achieved median Dice similarity coefficients ranging from 0.81 to 0.91 for different substructures



**Fig. 6 | Detailed information on the contouring time, Dice and HD improvements across different ROI and clinics. a, b** The percentage reduction in the average contouring time (indicated by the color bar) for phase one and two, respectively. **c, d** The absolute percentage increase in the Dice score and the decrease in HD between the manual and auto-edit ROI (from Fig. 4) for phase two. The size of

the markers in (a) and (b) represent the average manual contouring time, while the size of the markers in (c) and (d) represent the average Dice and HD between the manually contoured ROI and its consensus ROI. In all the figures, red color indicates improvement in auto-edit compared to manual contouring process. HD Hausdorff distance, ROI region-of-interest, LN lymph node, L left, R right.

on an independent test set. This segmentation model was implemented in the Contour+ auto-segmentation product by Mvision AI and the head-and-neck models had been deemed clinically acceptable in two independent single institutional studies<sup>7,34</sup>.

**Multicentre auto-segmentation study design.** In phase one of the study, each clinic first selected five anonymized head-and-neck planning CT scans according to the inclusion and exclusion criteria. After which, the radiation oncologist from each clinic would contour the 18 ROIs manually on all the five scans. 16 of the 18 ROIs were the bilateral parotid glands, submandibular glands, brachial plexus and different LN levels comprising of IB, II, III, IVA and IVB. The remaining two ROIs were the constrictor muscle and the LN IA. While clinical practice typically involves contouring LN levels without sub-level differentiation, the granularity employed in this study was important for two primary reasons: to enable sub-level-specific targeting<sup>35</sup> and to identify specific LN regions requiring further development in auto-segmentation accuracy<sup>36</sup>. There were other important OARs in head-and-neck such as brainstem and optic apparatuses, that were not included in this study. The 18 ROIs were selected by clinicians and Mvision, who agreed that these regions continue to pose a significant challenge for accurate contouring by most available auto-segmentation software. All the clinicians in each clinic were informed to contour the 18 ROIs based on the local institution contouring practices. The scans were sent to Mvision AI Oy (Helsinki,

Finland) to generate the auto-segmentations by DL. After a minimum two weeks, these auto-segmentations were returned to the respective clinic for the same clinician to review and edit the contours if required to achieve clinical quality. A minimum time delay was necessary to remove any bias associated with memory of the manual contours while editing the auto-segmentations. The aim of phase one was to study the auto-segmentation quality and the local contouring time saving with auto-segmentation in each institution.

In phase two of the study, two scans were selected randomly from each clinic. One scan would be used for manual contouring, while the other one would be used for editing of auto-segmented contours. The pooled seven CT scans which was meant for manual contouring were then circulated among the seven clinics. The pooling and exchange of the scans was to reduce any institutional scan-specific bias when evaluating the clinical benefit of the auto-segmentation. Each clinic then manually contoured the 18 ROIs for six CT scans from the other clinics. In addition, the auto-segmentations were also generated by Mvision for another set of seven scans and the same clinician from each clinic edited the auto-segmentations in all the six CT scans to achieve clinical quality as defined in their own clinic. The random selection and exchange of scans was facilitated by an independent party (J. Niemelä and G. Bolard from Mvision AI) who was not part of the participating hospitals to ensure that each clinic did not receive their own scan for manual contouring or editing. In order to further minimise the risk of potential recall bias, a different auto-segmented CT scan from the manually

contoured dataset was presented to each clinic. While this decision may diminish the statistical power to discern differences in contouring time or inter-observer contouring variability during phase two, it preempts the potential confounding effects of recall bias. In summary, the rationale for a two phases study design was to distinguish the clinical benefits assessed in a local context (phase one) and in a global context (phase two). In particular, through a series of scan exchanges, we aimed to quantify both the reduction in contouring time and inter-observer variability when using auto-segmentation, allowing for generalizable results.

### Contouring time saving in phase one and two

The time required to manually contour and to edit the auto-segmentation for each of the 18 ROIs were recorded in all the clinics in phase one and two. The time saving benefit of an auto-segmentation software was quantified by the difference between the time to edit the auto-segmentation and to manually segment the ROI from scratch. The time reduction of each ROI in all the 35 CT scans (five CT scans per clinic) scans in phase one were tested for statistical significance using the Wilcoxon signed-rank test. Similar statistical test was also used to test for significant time reduction in the ROI contouring in phase two (7 CT scans circulated among 7 clinics).

### Auto-segmentation performance in phase one

The contour similarity of the auto-segmentations versus manual contours and the edited versus manual contours were compared for each ROI in phase one. The aim was to quantify the performance of the auto-segmentation software with respect to the manual contour in each clinic and to quantify the difference between the manual and edited contours. The metrics for quantifying the contour similarity were Dice score and 95<sup>th</sup> percentile Hausdorff distance (referred to as HD in this study). Dice score is a volumetric overlap metric and it ranges from zero to one where zero and one represent no and perfect overlaps respectively<sup>37,38</sup>. 95<sup>th</sup> percentile HD<sup>39,40</sup> is a spatial distance-based metric, which is sensitive to boundary error<sup>39,41</sup> and is therefore not correlated to Dice score and plays a complementary role in this study<sup>37,42</sup>. It ranges from zero to infinity with zero indicating perfect overlap of the contours. These two metrics were chosen to give a holistic assessment of the contour agreement in terms of volumetric overlap and boundary error, and had been traditionally been used to quantify the quality of the semantic segmentation<sup>38,43,44</sup>. It is worth noting that there are other metrics such as the added path length (APL) and surface Dice score which were recently shown to be indicative of the contouring time saving<sup>37,45</sup>. However, since the contouring time was explicitly measured in this study, we decided to focus on the two most used metrics in segmentation task. Wilcoxon signed-rank test was used to test for significant difference of the metrics between the auto-segmentations and the edited contours for each ROI.

### Inter-observer variation in phase two

Inter-observer variation of a ROI measures the agreement between the contours delineated by different observers (or in this case a clinician) on a common CT scan. In phase two, the inter-observer variation between the seven clinics was compared for the manual contours and the edited contours for each ROI. The inter-observer variation of the manual and edited contours was quantified by two complementary approaches. The first involved calculating the pairwise contour similarity metrics (Dice and HD as defined above) between the seven manual contours and the consensus contour (calculated from the seven manual contours) for every ROI. The consensus contour was generated using the STAPLE (Simultaneous Truth And Performance Level Estimation) algorithm<sup>46,47</sup>. The same calculation was also performed for the seven edited contours and its consensus contour. The second approach was through the calculation of pairwise contour similarity metrics among the seven manual contours to yield a total of 21 calculations. Similar calculations were also performed among the edited contours. Large inter-observer variation was indicated by a poor contour similarity result in both approaches. A threshold Dice score of 0.80<sup>27</sup> was regarded as a good

overlap of agreement in the contours in the medical auto-segmentation task. There is currently no clear threshold for HD for an acceptable contour agreement and this served as a relative comparison in this study and to inform the reader the displacement deviation in the contours. Mann-Whitney U test was used to test for significant difference between the inter-observer variation between the manual contours and edited contours for each ROI.

### Statistical analysis

All the statistical tests were conducted using the *statsmodels v0.14.3* library in Python. In this study, a two-sided *P*-value < 0.05 was considered significant and Bonferroni correction was applied for multiple hypothesis testing.

### Data availability

The imaging and structure dataset used in this study cannot be shared publicly due to patient privacy and ethical restrictions. However, the results data generated from this study are available from the corresponding author upon reasonable request.

Received: 26 November 2024; Accepted: 9 April 2025;

Published online: 27 May 2025

### References

- Mukherjee, S., Vagha, S. & Gadkari, P. Navigating the Future: a comprehensive review of artificial intelligence applications in gastrointestinal cancer. *Cureus* **16**, e54467 (2024).
- Lotter, W. et al. Artificial intelligence in oncology: current landscape, challenges, and future directions. *Cancer Discov.* **14**, 711–726 (2024).
- Huynh, E. et al. Artificial intelligence in radiation oncology. *Nat. Rev. Clin. Oncol.* **17**, 771–781 (2020).
- Rong, Y. et al. NRG oncology assessment of artificial intelligence deep learning-based auto-segmentation for radiation therapy: current developments, clinical considerations, and future directions. *Int. J. Radiat. Oncol. Biol. Phys.* **119**, 261–280 (2024).
- Harrison, K. et al. Machine learning for auto-segmentation in radiotherapy planning. *Clin. Oncol. (R. Coll. Radio.)* **34**, 74–88 (2022).
- Cardenas, C. E., Yang, J., Anderson, B. M., Court, L. E. & Brock, K. B. Advances in auto-segmentation. *Semin Radiat. Oncol.* **29**, 185–197 (2019).
- Doolan, P. J. et al. A clinical evaluation of the performance of five commercial artificial intelligence contouring systems for radiotherapy. *Front Oncol.* **13**, 1213068 (2023).
- Ye, X. et al. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. *Nat. Commun.* **13**, 6137 (2022).
- Gibbons, E. et al. Clinical evaluation of deep learning and atlas-based auto-segmentation for critical organs at risk in radiation therapy. *J. Med Radiat. Sci.* **70**(Suppl 2), 15–25 (2023).
- Ayyalusamy, A. et al. Auto-segmentation of head and neck organs at risk in radiotherapy and its dependence on anatomic similarity. *Radiat. Oncol. J.* **37**, 134–142 (2019).
- Anand, A. et al. Study Design: Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning. *medRxiv*, 2021.2012.2007.21266421, <https://doi.org/10.1101/2021.12.07.21266421> (2021).
- Harari, P. M., Song, S. & Tomé, W. A. Emphasizing conformal avoidance versus target definition for IMRT planning in head-and-neck cancer. *Int. J. Radiat. Oncol., Biol., Phys.* **77**, 950–958 (2010).
- van der Veen, J., Gulyban, A., Willems, S., Maes, F. & Nuyts, S. Interobserver variability in organ at risk delineation in head and neck cancer. *Radiat. Oncol.* **16**, 120 (2021).
- Caissie, A. et al. Head and neck radiation therapy patterns of practice variability identified as a challenge to real-world big data: results from the learning from analysis of multicentre big data aggregation (LAMBDA) consortium. *Adv. Radiat. Oncol.* **8**, 100925 (2023).

15. Nielsen, C. P. et al. Interobserver variation in organs at risk contouring in head and neck cancer according to the DAHANCA guidelines. *Radiother. Oncol.* **197**, 110337 (2024).
16. Lucido, J. J. et al. Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning. *Front. Oncol.* <https://doi.org/10.3389/fonc.2023.1137803> (2023).
17. Stapleford, L. J. et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. *Int. J. Radiat. Oncol. Biol. Phys.* **77**, 959–966 (2010).
18. AlMBERG, S. S. et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. *Radiother. Oncol.* **173**, 62–68 (2022).
19. Singh, S., Singh, B. K. & Kumar, A. Multi-organ segmentation of organ-at-risk (OAR's) of head and neck site using ensemble learning technique. *Radiogr. (Lond.)* **30**, 673–680 (2024).
20. Strijbis, V. I. J. et al. Deep learning for automated elective lymph node level segmentation for head and neck cancer radiotherapy. *Cancers* **14**, 5501 (2022).
21. Samarasinghe, G. et al. Deep learning for segmentation in radiation therapy planning: a review. *J. Med. Imaging Radiat. Oncol.* **65**, 578–595 (2021).
22. Luan, S. et al. Accurate and robust auto-segmentation of head and neck organ-at-risks based on a novel CNN fine-tuning workflow. *J. Appl. Clin. Med Phys.* **25**, e14248 (2024).
23. Kiljunen, T. et al. A deep learning-based automated CT segmentation of prostate cancer anatomy for radiation therapy planning—a retrospective multicenter study. *Diagnostics (Basel)* <https://doi.org/10.3390/diagnostics10110959> (2020).
24. Udupa, J. K. et al. Combining natural and artificial intelligence for robust automatic anatomy segmentation: Application in neck and thorax auto-contouring. *Med Phys.* **49**, 7118–7149 (2022).
25. Kosmin, M. et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother. Oncol.* **135**, 130–140 (2019).
26. Ginn, J. S. et al. A clinical and time savings evaluation of a deep learning automatic contouring algorithm. *Med Dosim.* **48**, 55–60 (2023).
27. Brock, K. K., Mutic, S., McNutt, T. R., Li, H. & Kessler, M. L. Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132. *Med Phys.* **44**, e43–e76 (2017).
28. Mir, R. et al. Organ at risk delineation for radiation therapy clinical trials: Global Harmonization Group consensus guidelines. *Radiother. Oncol.* **150**, 30–39 (2020).
29. Brouwer, C. L. et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother. Oncol.* **117**, 83–90 (2015).
30. Melerowitz, L. et al. Design and evaluation of a deep learning-based automatic segmentation of maxillary and mandibular substructures using a 3D U-Net. *Clin. Transl. Radiat. Oncol.* **47**, 100780 (2024).
31. Kong, F. M. et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. *Int. J. Radiat. Oncol. Biol. Phys.* **81**, 1442–1457 (2011).
32. Scoccianti, S. et al. Organs at risk in the brain and their dose-constraints in adults and in children: a radiation oncologist's guide for delineation in everyday practice. *Radiother. Oncol.* **114**, 230–238 (2015).
33. Gregoire, V. et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. *Radiother. Oncol.* **110**, 172–181 (2014).
34. Strolin, S. et al. How smart is artificial intelligence in organs delineation? Testing a CE and FDA-approved Deep-Learning tool using multiple expert contours delineated on planning CT images. *Front Oncol.* **13**, 1089807 (2023).
35. Biau, J. et al. Selection of lymph node target volumes for definitive head and neck radiation therapy: a 2019 Update. *Radiother. Oncol.* **134**, 1–9 (2019).
36. Luo, X. et al. A multicenter dataset for lymph node clinical target volume delineation of nasopharyngeal carcinoma. *Sci. Data* **11**, 1085 (2024).
37. Sherer, M. V. et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother. Oncol.* **160**, 185–191 (2021).
38. Taha, A. A. & Hanbury, A. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Med Imaging* **15**, 29 (2015).
39. Sim, D. G., Kwon, O. K. & Park, R. H. Object matching algorithms using robust Hausdorff distance measures. *IEEE Trans. Image Process* **8**, 425–429 (1999).
40. Christiaens, M. et al. Quality assurance of radiotherapy in the ongoing EORTC 1219-DAHANCA-29 trial for HPV/p16 negative squamous cell carcinoma of the head and neck: Results of the benchmark case procedure. *Radiother. Oncol.* **123**, 424–430 (2017).
41. Sharp, G. et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys.* **41**, 050902 (2014).
42. Hanna, G. G., Hounsell, A. R. & O'Sullivan, J. M. Geometrical analysis of radiotherapy target volume delineation: a systematic review of reported comparison methods. *Clin. Oncol.* **22**, 515–525 (2010).
43. AlBadawy, E. A., Saha, A. & Mazurowski, M. A. Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med Phys.* **45**, 1150–1158 (2018).
44. Maier-Hein, L. et al. Metrics reloaded: recommendations for image analysis validation. *Nat. Methods* **21**, 195–212 (2024).
45. Vaassen, F. et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Phys. Imaging Radiat. Oncol.* **13**, 1–6 (2020).
46. Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Med. Imaging* **23**, 903–921 (2004).
47. Hwee, J. et al. Technology assessment of automated atlas based segmentation in prostate bed contouring. *Radiat. Oncol.* **6**, 110 (2011).

**Author contributions**

Study conception and design: E.P.P Pang, H.Q. Tan, M.L.K. Chua Data acquisition and analysis: All authors Data interpretation: E.P.P Pang, H.Q. Tan, M.L.K. Chua Statistical analyses: H.Q. Tan, J. Niemelä, G. Bolard Drafting of manuscript: E.P.P Pang, H.Q. Tan Approval of final manuscript: All authors.

**Competing interests**

The authors declare that J. Niemelä is one of the founders and shareholders of Mvision AI Oy. All other authors declare no competing interests.

**Additional information**

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41746-025-01624-z>.

**Correspondence** and requests for materials should be addressed to Hong Qi Tan.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025