nature portfolio

Peer Review File

A prospective multicenter trial of deep learning auto-segmentation for organs at risk in thoracic radiotherapy

Corresponding Author: Professor Zhiyong Yuan

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
General remarks
The manuscript describes the results of a clinical trial to evaluate the clinical performance of a deep learning-based autocontouring (AI) method for delineation of organs at risk in thoracic and breast cancer radiotherapy. The study design is very strong because not only the performance of the AI method was considered but also the quality of the contours after manual adjustment by a radiation oncologists was assessed. Furthermore, 500 patients from 5 different hospitals were included and alongside the AI-generated contours, two human delineators provided both manual and AI-adjusted contours. All those contours were compared with ground truth contours, generated by three independent radiation oncologists that did not participate in the generation of the evaluated contours. This setup enables a robust comparison between fully manual and AI-assisted workflows.
However, some limitations exist in the study and the current manuscript that need to be solved to make the manuscript suitable for publication:

1. The segmentation performance is evaluated using only volume-based metrics. However, surface- and distance-based metrics (e.g., surface DSC, Hausdorff distance) have proven valuable in literature, particularly in addressing the limitations of purely volumetric metrics such as their dependency on structure size. Including these metrics is essential for a thorough evaluation of the segmentations.
2. The manuscript references the iCurveE model as if it were well-known, but this is not a standard or widely recognized model and information on this model is lacking online. There is a lack of citation, and important details on the model architecture, training, validation, and testing pipeline are missing. Key aspects such as CT resolution, specific data augmentation parameters (e.g., rotation degrees and probabilities), training software, hyperparameter tuning strategies, model selection criteria, and the handling of missing structures or ground truth availability is missing.
3. Although the analysis is thorough and many results are available, the information in the main results on model performance is limited. For example, a visualisation of AI-only performance and differences between centres is lacking in figure 2a. Although many results are available, many are left to the reader's own interpretation by merely mentioning them in the supplementary material.
4. The study focuses on OARs that are relatively straightforward to segment and have already demonstrated high performance in previous auto-segmentation studies (e.g., lung, liver, whole heart). Including more anatomically complex structures, such as heart substructures, kidneys, lymph node levels, breast, chest wall, brachial plexus etc. could have provided greater insight into the limitations and biases of auto-segmentation, especially in clinically challenging scenarios.

Detailed comments
1. Line 80: you mention ethical concerns, but also legal concerns exist. Please add information on legal requirements for human oversight for AI applications in medicine (e.g. European AI Act).
2. Line 101: It is unclear which structures were (if any at all) were used in clinic. To make it more confusing, the authors mention that the study aimed to "simulate real-world clinical scenarios" was it all a simulation or actual clinical workflow?
3. Results line 116: now the results section focused on manual and AI assisted contouring. It would be interesting to also add the AI contours results here. To be able to decide whether manual adjustments are needed.
4. Results: to get an idea of the clinical impact of the differences between the manual, AI and AI assisted contours also the impact on the relevant DVH parameters (those that are used for treatment plan optimization) should be added. DVH parameters of OARs far away from the target volume will be less sensitive to small contour differences.

5. Discussion line 274-276: the absence of distance-based metrics (like the Hausdorff distance) and the surface dice is a big limitation of this study. Volume DSC values depend on the volume of the contoured structure and are therefore not very sensitive to contour differences in case of large structures like the lungs. Therefore, results based on a distance-based metrics and the surface dice need to be added to this study to be able to draw reliable conclusions on the contouring accuracy of the different methods. Those analyses can be added afterwards and do not need an adjustment of the trial design or collection of additional data.
6. Materials and methods, study design: The process of contour assignment is ambiguous. It remains unclear how many contours were seen by individual delineators and if each contour got assigned one person from 2 different centers.
7. Materials and methods, study design: please add information on the contouring process. Were detailed contouring guidelines available for all OARs used in the study with detailed definitions of all borders of the OARs? Were those same guidelines used in all structures used in the study: the ground truth structures, the structures used for model training and the structures used for the study?
8. Materials and methods, Computed tomography, line 322: both non-contrast and contrast-enhanced scans were accepted. Please add information on the numbers of scans with and without contrast in supplementary table 1.
9. Materials and methods, Computed tomography and supplementary table 1: it is striking that 75% of all CTs had a slice thickness of 5 mm. This is very large for accurate contouring of small structures. This limitation of the study should be added to the discussion.
10. Materials and methods, Computed tomography and supplementary table 1: add information on the in-slice resolution of the CT-scans.
11. Materials and methods, Model information. Was the model trained on other patients than those used in the analysis? Is Sun Yat-sen University (SUH) Cancer center another hospital than The Fifth Affiliated Hospital of Sun Yat-sen University (SUH)? If this is the same hospital, than please clarify whether the patients used in model training are different from the ones used in the analysis.
12. Materials and methods, evaluation metrics: please add at least a distance-based metric (like Hausdorff distance) and the surface dice as evaluation metrics, to make the results less biased, since the volumetric DSC depends on the volume of the structures.
13. Results: The analysis does not include any comparison between contrast-enhanced and non-contrast CT scans, which is particularly important for OARs such as the aorta, whose appearance and delineation can be significantly affected by the presence of contrast.
14. Results, first paragraph: Please double-check the numerical data in the first paragraph of the results section. The numbers do not appear to align correctly (e.g., 500 ≠ 146 + 35 + 320), and there is inconsistency in the number of manual and AI-assisted sets reported (e.g., 993 vs. 992 manual sets, and 497 AI-generated sets, which is not half of 993).
15. Figure 1: While you explain the discrepancy in patient numbers in the figure caption, please also clarify this in the main text for consistency. Additionally, consider standardising the terminology, 'manual re-contouring' and 'manual delineation' refer to the same process but are currently used interchangeably. Consistent phrasing will improve clarity. Please clarify what you mean with “insufficient CT scan or model error” and also mention that this exclusion was performed in the main text.
16. Figure 2b: The equation is written as M-Aa/M (i.e. M - (Aa/M)), but it seems you likely intended to use (M - Aa)/M. Please verify and correct the expression to ensure it reflects the intended calculation.
17. Figure 2a, please add also the results after AI segmentation in this figure.
18. Figure 2c: please write CT-slices instead of CT-layers.
19. Figure 3a: please add more information on the Likert scale used in this assessment in the materials and methods section. Where the only possible choices: strongly satisfied, satisfied of neutral? Or was also a more negative choice possible? Was this assessment performed for every AI segmentation that was modified?
20. How do you explain that the manual adjustment of an AI segmentation that was strongly satisfied/satisfied took more time than for an AI segmentation that was neutral (figure 3f)? I would expect that this would be the other way around.
21. In Figure 4b and 4f, the use of circles to indicate DSC improvement is visually unclear. Consider an alternative, more intuitive visual representation for this data.
22. Figure 4b: why did you distinguish between previous and next centers? I would just present the differences in DSC between two centers independent of the order of segmentation.
23. Figure 4f: please explain the abbreviations of the OARs in the caption.
24. In Figure 5b: it only makes sense that OARs with a lower DSC have more DSC improvements, because their scores were lower to begin with. It would be more interesting to see DSC of the AI as a function of manual DSC and DSC of adjusted AI as a function of M-DSC to see whether more difficult structures are also more difficult for the network.

Supplementary data:
25. Table 1 and figure 2 of the supplementary data contain very relevant information and should preferably added to the main manuscript. Figure 2a of the main manuscript could be adjusted to add also the AI data. Than no additional figure is necessary. To see the manual, AI and AI-assisted results in one figure would be very insightful.

26. The supplementary data contain many analyses and detailed tables, which are not all cited in the main manuscript. Now it is not easy for the reader to draw conclusions from all this information. Some adjustments are necessary:
a. present the most relevant information on contouring accuracy quantified in DSC or VRI in figures (like figure 2a in the main paper). Figures can be interpreted much easier than those busy tables. I suggest that you at least make figures for supplementary tables 3,6, 15 and 21. Also graphs for the contouring times makes a fast interpretation easier than tables.
b. Add some text in the supplemental materials where you summarize the main findings and conclusions from the tables.
c. Consider removing some less relevant supplemental materials.

27. Add a table and graph for the comparison of DSC in CT-scans with and without contrast.

28. In all tables concerning contouring time the unit of time is missing: please add this. I guess it should be minutes.
29. Supplementary figure 3: it seems that the physician did not want to take enough time anymore to contour the lungs manually at the end of the study. Maybe because using the AI contours was much easier? Please add this in your discussion.
30. Supplementary Method1: data preprocessing: what was the resolution and slice thickness of the used CTs? What was the voxel size of the images after resampling?
31. Add more information on the deep learning autocontouring model used in this study in Method 2. What was the performance of the model in the test set? What kind of post-processing was used. Etc. Can the code of this model be shared?
32. Supplementary table 33: what kind of AEs and SAEs during the CT imaging process are you referring to? Were the contours used in the clinic?

Reviewer #2

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Reviewer #3

(Remarks to the Author)
The study addresses a timely and clinically relevant topic, and I commend the authors for conducting a prospective, multicenter evaluation of an AI-based segmentation tool in routine radiotherapy workflows. However, the authors failed to leverage this opportunity to demonstrate the potential impact of AI in clinical practice.

Strengths:
1. Prospective multicenter design: The use of five clinical centers and a large, diverse group of physicians enhances the generalizability and clinical relevance of the findings.

2. Robust annotation and evaluation protocol: The study is meticulously designed, involving expert-defined ground truth annotations and comparative evaluation across manual, AI, and AI-assisted methods.

3. Comprehensive performance assessment: Multiple organs at risk (OARs) were assessed using key performance metrics (e.g., Dice similarity coefficient, time efficiency, and volumetric revision index), providing a thorough evaluation of the segmentation tool.

Major Comments of Weaknesses
1. Description of the deep learning model
The segmentation model (iCurveE) appears to be a commercial tool, but the manuscript lacks detailed description of the underlying architecture, training datasets, or performance benchmarking against widely adopted open-source models (e.g., nnUNet). Inclusion of these details would allow readers to assess the novelty and generalizability of the model.

2. Clinical relevance beyond segmentation accuracy
While the study convincingly demonstrates improvements in segmentation efficiency and accuracy, it does not evaluate the downstream impact on radiation treatment planning or patient outcomes. Metrics such as changes in dose-volume histogram parameters or plan quality would significantly enhance the clinical relevance of the findings.

3. Justification for trial design
The rationale for using a cross-over trial design is not clearly articulated, especially given that patients were treated based on ground truth contours. Similarly, the purpose of cross-center contouring is unclear—whether it is intended to assess inter-physician variability, model generalizability, or end-user learning effects.

4. Evaluation metrics
While Dice coefficient is widely used, it may not be sufficiently sensitive for large or irregularly shaped organs. Supplementing with additional metrics such as Hausdorff distance, mean surface distance, or dosimetric impact (e.g., changes in Dmax, Dmean) would provide a more comprehensive evaluation.

Below is the review comments from Early Career Researcher who co-reviewed this paper:

The manuscript still needs a clearer justification for developing a 2-D Res-SE-Net (“iCurveE”) rather than leveraging the many open-source or commercial solutions that already delineate thoracic OARs. To make the technical contribution unambiguous, please add a evaluation with three readily available comparators:
• Open-source baseline (trained from scratch). Train a standard nnU-Net on the identical multicentre cohort and report Dice, HD95 and inference time. This will show whether your network design itself, not merely the dataset, drives the reported gains.
• Pre-trained, publicly released tool. Run Total Segmentator (which already contours all but the whole humerus vs. humeral

head) on the test set. A quantitative comparison will reveal whether fine-tuning an existing model could achieve comparable accuracy and time-savings without developing a new architecture.
• Commercial FDA-cleared TPS solutions. Where licences permit, include at least one widely used system that offers integrated OAR contouring such as MIM contour Protege, RayStation, or Eclipse. Such a benchmark would highlight your results in clinical application and underscore any added value for busy radiotherapy departments.

Please incorporate these experiments or, if infeasible, provide a detailed rationale (with citations) for omitting them. Clarifying the performance gap between iCurveE and these baselines will substantially strengthen the paper's claim of technical and clinical novelty.

Dosimetric Evaluation: Even though the authors explicitly acknowledge the lack of dosimetric comparison, While the geometric improvements reported for AI-assisted contours are highlighted, the manuscript stops short of demonstrating their clinical relevance. Since the proposed pipeline is intended for "treatment planning", please include a dosimetric comparison between plans generated with (i) ground-truth contours, (ii) manual physician contours, and (iii) AI/AI-assisted contours. At minimum, report standard dose-volume metrics, D_mean, D_max, D_min, D1, D50, and volume-based indices such as V1, V10, V20, V50, for the eleven OARs. A subset analysis on 20–30 representative cases would already clarify whether a 4–5 percentage-point Dice increase (e.g., 0.85 → 0.90) yields a tangible reduction in dose to critical structures.

Distance-based accuracy: The authors acknowledge that HD95 or surface distance were excluded because the decision pre-dated NRG Oncology guidance, leaving only volume metrics (DSC, VRI) . For small, serial OARs (e.g. esophagus, brachial plexus) border errors can be clinically more important than volumetric overlap. A limited HD95 analysis on 20–30 test cases would strengthen the claim that AI assistance improves clinically important accurac

Ablation Study for training cohort: The model was trained on 827 scans from only two Chinese hospitals , yet tested on five. An ablation in which those two centerws are withheld from evaluation (true external test) would quantify domain-shift robustness.

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
My compliments for your clear and extensive reply to our comments and for the efforts you have taken to improve the manuscript based. The quality of the manuscript has improved significantly and you have replied adequately to all our comments.

We appreciate your efforts in improving the visualisations for model comparisons. Especially supplental figure 3 with the boxplot representation gives a very clear overview of the differences between the different models. However, in Supplementary Figure 2, the chosen heatmap representation still makes it challenging to directly compare the iCurve model with nnU-Net and TotalSegmentator. We suggest you to combine Supplementary Figures 2 and 3 and show the results for both comparisons as boxplots.

Similarly, for Figures 4c–i, 6a–b, and Supplementary Figures 5, 6, 7, 8, 9, 10, 11, and 17, we suggest you to include the numerical values inside the boxes of the heatmaps to improve clarity and interpretability. Currently, while the heatmaps provide a clear visual overview, they do not fully convey the quantitative differences between models.

Reviewer #3

(Remarks to the Author)
The authors have sufficiently addressed my concerns, no more questions.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit https://creativecommons.org/licenses/by/4.0/

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

General remarks

The manuscript describes the results of a clinical trial to evaluate the clinical performance of a deep learning-based autocontouring (AI) method for delineation of organs at risk in thoracic and breast cancer radiotherapy. The study design is very strong because not only the performance of the AI method was considered but also the quality of the contours after manual adjustment by a radiation oncologists was assessed. Furthermore, 500 patients from 5 different hospitals were included and alongside the AI-generated contours, two human delineators provided both manual and AI-adjusted contours. All those contours were compared with ground truth contours, generated by three independent radiation oncologists that did not participate in the generation of the evaluated contours. This setup enables a robust comparison between fully manual and AI-assisted workflows. However, some limitations exist in the study and the current manuscript that need to be solved to make the manuscript suitable for publication:

Response:

We sincerely thank you for your meticulous review and constructive comments on our manuscript. We are particularly grateful for your recognition of our work and agree that the points you have raised are crucial for improving the manuscript. In response, we have thoroughly revised the manuscript to address these points, and our detailed, point-by-point responses are provided below.

Major Comment 1: **The segmentation performance is evaluated using only volume-based metrics. However, surface- and distance-based metrics (e.g., surface DSC, Hausdorff distance) have proven valuable in literature, particularly in addressing the limitations of purely volumetric metrics such as their dependency on structure size. Including these metrics is essential for a thorough evaluation of the segmentations.**

Response:

Thank you for your valuable suggestion to incorporate distance- and surface-based metrics. To address this point, we have meticulously re-analyzed all the **27,043** OARs from the **500** patients in our trial. We calculated both the 95th percentile Hausdorff Distance **(HD95)** and the Surface Dice Similarity Coefficient **(sDSC)** for each delineation and performed the corresponding statistical comparisons.

These new findings further validated our original conclusions: the iCurveE-based AI-assisted delineation **remained** significantly superior to manual delineation across all 11 OARs based on these new metrics.

Accordingly, we have extensively updated the manuscript to integrate these results, including revisions to the **main text, Figures (e.g., Figs. 2–6), and Supplementary Materials (e.g.,**

Supplementary Figs. 4–16). We believe these additions substantially enhance the robustness of our evaluation.

Major Comment 2: **The manuscript references the iCurveE model as if it were well-known, but this is not a standard or widely recognized model and information on this model is lacking online. There is a lack of citation, and important details on the model architecture, training, validation, and testing pipeline are missing. Key aspects such as CT resolution, specific data augmentation parameters (e.g., rotation degrees and probabilities), training software, hyperparameter tuning strategies, model selection criteria, and the handling of missing structures or ground truth availability is missing.**

Response:

As you correctly pointed out, while our iCurveE model has received **Class III** clinical approval from China's National Medical Products Administration (NMPA) (No. 20253210066), it currently lacks prior academic reporting. Accordingly, we have made the following additions:

1. We have added the corresponding details of the model's architecture and development in **Supplementary Methods 1–2 and Supplementary Fig. 22**, as per your guidance.
2. We have included the imaging information for model development (e.g., contrast status, resolution, and slice thickness) in **Supplementary Table 5.**
3. We have presented the model's performance on the training, validation, and test sets (based on volumetric DSC and HD95) in **Supplementary Table 6.**
4. We have benchmarked iCurveE against an nnU-Net trained from scratch (using the same training cohort) and the open-source TotalSegmentator, presenting the results in a new section titled **"Performance benchmark of the iCurveE model" and in Supplementary Figs. 2–3.**

Thank you once again for this constructive feedback. We believe that providing these model details has substantially enhanced the transparency and rigor of our manuscript.

Major Comment 3: **Although the analysis is thorough and many results are available, the information in the main results on model performance is limited. For example, a visualisation of AI-only performance and differences between centres is lacking in figure 2a. Although many results are available, many are left to the reader's own interpretation by merely mentioning them in the supplementary material.**

Response:

Thank you for this constructive feedback. We agree that the manuscript would be strengthened by a more direct and comprehensive visualization of the AI model's standalone performance.

To address this point, we have made the following major revisions based on a multi-metric evaluation (e.g., vDSC, HD95, sDSC, VRI):

1. In **Fig. 2a–b and Supplementary Figs. 4 and 13,** we have added a comparison of the AI delineation performance with that of the manual and AI-assisted delineations.

2. In **Fig. 4a–b and Supplementary Fig. 15,** we have added a new analysis that visualizes the AI-only performance across the five centers.
3. We have also added performance comparisons between manual, AI-only, and AI-assisted methods to various subgroup analyses throughout the **Supplementary Materials** (e.g., Supplementary Figs. 8–16).

Furthermore, we have also updated the main text with detailed descriptions of these findings. We believe these revisions significantly enhance the intuitive presentation of our model's performance.

Major Comment 4: The study focuses on OARs that are relatively straightforward to segment and have already demonstrated high performance in previous auto-segmentation studies (e.g., lung, liver, whole heart). Including more anatomically complex structures, such as heart substructures, kidneys, lymph node levels, breast, chest wall, brachial plexus etc. could have provided greater insight into the limitations and biases of auto-segmentation, especially in clinically challenging scenarios.

Response:

Thank you for your insightful comment. We fully agree that including more complex structures is of great clinical significance for exploring the boundaries and potential of radiotherapy auto-segmentation models.

However, this study represents the largest end-user testing of a radiotherapy auto-segmentation model to date, involving 500 patients, 27,043 OARs, and 37 physicians from 5 centers. Therefore, to supplement this work with manual, AI, AI-assisted, and ground truth delineations for complex substructures would require us to re-execute the entire study, which would constitute a new significant trial.

While these thoracic OARs are conventional in previous studies, we benchmarked our model (iCurveE) against a pre-trained TotalSegmentator and an nnU-Net trained from scratch to quantify its contribution in this competitive domain. In this context, our model still demonstrated better performance **(Supplementary Figs. 2–3).**

Nevertheless, we strongly agree with your point and have accordingly incorporated it as a **limitation** in the Discussion section of our revised manuscript.

Detailed comments

1. Line 80: you mention ethical concerns, but also legal concerns exist. Please add information on legal requirements for human oversight for AI applications in medicine (e.g. European AI Act).

Response:

Thank you for this valuable suggestion. We agree that highlighting the legal mandate for human oversight reinforces the importance of this study. Therefore, we have revised the **Introduction** accordingly and added three recent citations (2024-2025) to the original reference to underscore this

point:

"*Nevertheless, given the "black-box" nature of DL algorithms and corresponding ethical concerns, human oversight of AI-generated outputs is mandated by legal frameworks, such as the European Union Artificial Intelligence Act (EU AI Act), and reinforced by professional consensus.*[1-4]"

Thank you again for helping us improve the manuscript in this important aspect.

---

2. Line 101: It is unclear which structures were (if any at all) were used in clinic. To make it more confusing, the authors mention that the study aimed to "simulate real-world clinical scenarios" was it all a simulation or actual clinical workflow?

Response:

We sincerely thank you for your insightful comment and agree that our original phrasing, particularly the terms "simulate" and "real-world," were ambiguous.

To further clarify, this study was an observational clinical trial (NCT05787522). Although all enrolled CT scans were subsequently used to generate the clinical radiotherapy plans, the protocol **did not mandate** the clinical use of any specific contour set (manual, AI, AI-assisted, or ground truth). This design ensured that the research workflow **would not interfere with timely patient care** for these prospectively enrolled patients after their CT simulation, leaving the final decision on contouring to the treating physician. Specifically, the required washout period and the establishment of a multi-expert ground truth may delay the availability of the research contours.

Accordingly, we have (1) revised this sentence in **Introduction** as follows: "*This prospective multicenter study aimed to comprehensively compare the performance and clinical applicability of the DL model (iCurveE)-based AI-assisted delineation with traditional manual contouring for thoracic OARs, through large-scale end-user testing by physicians of varying expertise on planning CTs from heterogeneous patients with cancer.*" and (2) clarified the principles of clinical application for the research contours in the **Study design** subsection of the Methods.

---

3. Results line 116: now the results section focused on manual and AI assisted contouring. It would be interesting to also add the AI contours results here. To be able to decide whether manual adjustments are needed.

Response:

Thank you for your valuable suggestion. Although we briefly described the performance of the standalone AI model in our initial manuscript (lines 121-123), we realize that this description was likely not comprehensive or intuitive enough.

Accordingly, we have made two key revisions to address this. First, we have updated the relevant main figures (**Figs. 2a–b and 4a–b**) and Supplementary Materials (**e.g., Supplementary Figs. 2, 11, 14, and 15**) to provide a multi-metric comparison across all three methods (manual, AI, and AI-assisted). Second, we have revised the corresponding text in the **Results section** to more

comprehensively describe the standalone AI's performance and its comparison with the other contouring methods.

---

4. Results: to get an idea of the clinical impact of the differences between the manual, AI and AI assisted contours also the impact on the relevant DVH parameters (those that are used for treatment plan optimization) should be added. DVH parameters of OARs far away from the target volume will be less sensitive to small contour differences.

Response:

Thank you for this clinically crucial suggestion. We have now performed a dosimetric evaluation on 40 randomly selected patients. We calculated dosimetric deviations by comparing **38 DVH parameters** for each contouring method (manual, AI, and AI-assisted) against those derived from the ground-truth contours, using the fixed dose distribution from the clinical treatment plans. The methodology has been detailed in the "**Dosimetric Validation**" subsection of the Methods.

Our results show the mean dosimetric deviations were significantly lower for AI-assisted (4.2%) and AI-only (5.8%) delineations compared to manual delineation (11.1%). The AI and AI-assisted methods significantly reduced deviations in 13 (34.2%) and 23 (60.5%) of the DVH parameters, respectively **(Fig. 6a–b)**. Furthermore, we observed a significant correlation between higher geometric accuracy and smaller dosimetric deviations across all cancer types **(Fig. 6c–f)**. These results are now detailed in the Results subsection **"Reduced dosimetric deviations with AI-assisted delineation."**

As you pointed out, dosimetric errors are indeed more sensitive to contour inaccuracies near the target volume than to overall geometric accuracy, which aligns with our previous dosimetric studies.[5] Accordingly, in the **limitations section of the Discussion**, we have now elaborated on clinically relevant non-geometric factors that influence dosimetric impact and acknowledged the potential shortcomings of our preliminary, direct dosimetric evaluation.

Thank you again for this constructive feedback, which has markedly enhanced the comprehensiveness and clinical relevance of our study.

---

5. Discussion line 274-276: the absence of distance-based metrics (like the Hausdorff distance) and the surface dice is a big limitation of this study. Volume DSC values depend on the volume of the contoured structure and are therefore not very sensitive to contour differences in case of large structures like the lungs. Therefore, results based on a distance-based metrics and the surface dice need to be added to this study to be able to draw reliable conclusions on the contouring accuracy of the different methods. Those analyses can be added afterwards and do not need an adjustment of the trial design or collection of additional data.

Response:

Thank you again for this valuable suggestion. We have addressed this point in detail in our response to Major Comment 1 and believe this limitation has now been fully resolved.

6. Materials and methods, study design: The process of contour assignment is ambiguous. It remains unclear how many contours were seen by individual delineators and if each contour got assigned one person from 2 different centers.

Response：

Thank you for highlighting the need for more clarity on our contour assignment process.

Accordingly, we have detailed the number of manual and AI-assisted delineation cases for each physician in **Supplementary Table 3** and clarified that each patient's CT images were assigned to one physician from each of the two adjacent centers.

To ensure this process is transparent, we have revised the **Study Design** section to read as follows: *"…CT images acquired from each center were also circulated to the next center for manual and AI-assisted delineations (e.g., Center 1 to Center 2, ..., Center 5 to Center 1), thereby creating a dataset of shared images,* ***with each image set delineated by one physician from each of the two adjacent centers.***

Consequently, e*ach image set yielded two manual delineations and two AI-assisted delineations, forming pairs for comparisons: (1) one pair from a physician at the recruiting center for intra-physician paired comparison between manual and AI-assisted methods; and (2) one pair from a physician at the next center for inter-center and inter-physician comparisons (Fig. 1 and Supplementary Fig. 1). Additionally, a single, identical AI delineation set was generated for each image."*

---

7. Materials and methods, study design: please add information on the contouring process. Were detailed contouring guidelines available for all OARs used in the study with detailed definitions of all borders of the OARs? Were those same guidelines used in all structures used in the study: the ground truth structures, the structures used for model training and the structures used for the study?

Response：

Thank you for this important question regarding the standardization of our contouring process. To further clarify, we have added an **"OAR contouring standards" section** in the Materials and Methods and have provided the following explanation:

"Among the 11 OARs, the bilateral lungs, heart, esophagus, aorta, spinal cord, trachea, and bronchial tree were delineated according to the Radiation Therapy Oncology Group (RTOG) 1106 atlas,[6] *while the liver followed the RTOG consensus guidelines for upper abdominal normal organs.*[7] *The bilateral humeral heads were delineated as the bony structure from their apex to the surgical neck under a bone window, following standard radiological anatomy.*[8] *This guideline was applied to all delineations throughout the entire clinical trial."*

---

8. Materials and methods, Computed tomography, line 322: both non-contrast and contrast-enhanced scans were accepted. Please add information on the numbers of scans with and

without contrast in supplementary table 1.

Response：

Thank you for your valuable suggestion. Accordingly, we have reviewed the information for the 500 image sets and replaced the original Supplementary Table 1 with a new **Table 1** in the main manuscript to better highlight these details. The cohort includes 254 (50.8%) contrast-enhanced (CE) CT scans and 246 (49.2%) non-contrast (NC) CT scans.

Furthermore, we performed a multi-metric comparison of delineation performance on CE-CT and NC-CT scans, with the results presented in **Supplementary Figure 10.**

---

9. Materials and methods, Computed tomography and supplementary table 1: it is striking that 75% of all CTs had a slice thickness of 5 mm. This is very large for accurate contouring of small structures. This limitation of the study should be added to the discussion.

Response：

Thank you for your perceptive observation. To investigate this, we performed a sub-analysis **comparing** delineation performance on CTs with slice thicknesses of **<5 mm versus 5 mm** using multiple metrics. Our results confirmed that for smaller structures like the **humeral heads**, the 5 mm slice thickness was associated with a lower performance for the iCurveE model ("**Impact of clinical and imaging heterogeneity" subsection and Supplementary Fig. 11**).

However, the AI delineation accuracy for the humeral heads still surpassed that of manual delineation (Fig. 2a-b) and the nnU-Net model trained from scratch (Supplementary Fig. 2). Moreover, given the prospective collection of simulation CTs in this trial, the distribution of slice thicknesses **partially reflects real-world clinical practice**, potentially validating the model's accuracy and robustness when faced with challenges in clinical imaging quality.

Therefore, in accordance with your valuable suggestion, we have added a relevant discussion of this limitation to the Discussion section:

"First, a significant portion (75.6%) of our study's CT scans had a 5-mm slice thickness. This large thickness may introduce "partial volume effects", [9,10] potentially impacting model performance for small-volume structures like the humeral heads (Supplementary Fig. 11). However, given the prospective design of this study, this data distribution partially reflects real-world clinical practice, thereby validating the model's robustness when faced with challenges in clinical imaging quality."

---

10. Materials and methods, Computed tomography and supplementary table 1: add information on the in-slice resolution of the CT-scans.

Response：

Thank you for this helpful suggestion. To address this, we have added the in-slice resolution information for the 500-scan study cohort and the 827-scan model development cohort to **Table 1** and **Supplementary Table 5**, respectively.

---

11. Materials and methods, Model information. Was the model trained on other patients than those used in the analysis? Is Sun Yat-sen University (SUH) Cancer center another hospital than The Fifth Affiliated Hospital of Sun Yat-sen University (SUH)? If this is the same hospital, please clarify whether the patients used in model training are different from the ones used in the analysis.

Response：

Thank you for highlighting this potential point of confusion. To clarify, the model development and study cohort are entirely separate and were sourced from different centers. As stated in the **"Model information"** section, "The model was developed on a retrospective dataset of 827 planning chest CTs … **with no center, patient, or delineator overlap** with the 5-center cohort of this study." The prospective nature of this trial inherently ensures this separation.

In response to your valuable question, we have now abbreviated The Fifth Affiliated Hospital of Sun Yat-sen University (the trial center) and Sun Yat-sen University Cancer Center (the source of training images) as **FHSU and SYSUCC**, respectively. We believe these revisions have adequately addressed your concern.

---

12. Materials and methods, evaluation metrics: please add at least a distance-based metric (like Hausdorff distance) and the surface dice as evaluation metrics, to make the results less biased, since the volumetric DSC depends on the volume of the structures.

Response：

Thank you again for this valuable suggestion. Accordingly, we have now incorporated both HD95 and the Surface Dice (sDSC) in this section. Moreover, we have described their calculation methods in Supplementary Table 7 and integrated the results throughout the relevant sections.

---

13. Results: The analysis does not include any comparison between contrast-enhanced and non-contrast CT scans, which is particularly important for OARs such as the aorta, whose appearance and delineation can be significantly affected by the presence of contrast.

Response：

Thank you very much for this constructive feedback. In response, we have performed a detailed comparison of delineation performance on contrast-enhanced versus non-contrast CTs in manual, AI, and AI-assisted delineation. The results are now presented in **Supplementary Fig. 10** and described in a new section titled **"Impact of clinical and imaging heterogeneity"**.

The results showed that for the well-vascularized OARs (heart, aorta, and liver), delineation on contrast-enhanced CTs **consistently outperformed** that on non-contrast CTs across all three methods (manual, AI, and AI-assisted), as measured by vDSC, HD95, and sDSC.

---

14. Results, first paragraph: Please double-check the numerical data in the first paragraph of the results section. The numbers do not appear to align correctly (e.g., 500 ≠ 146 + 35 + 320), and there is inconsistency in the number of manual and AI-assisted sets reported (e.g., 993 vs.

992 manual sets, and 497 AI-generated sets, which is not half of 993)

Response：

Thank you for your meticulous review. We sincerely apologize for any confusion regarding the number of patient cohort and delineation sets.

1. Regarding the patient numbers (500 vs. 146+35+320): As you also kindly suggested in a subsequent comment, we have now revised the sentence in the Results section to read: "*...including 146 with LC..., 35 with EC, and 320 with BC (Table 1),* ***while 1 patient with concurrent lung and breast cancer was counted in both cohorts.***"

2. **Regarding the delineation set numbers (993 vs. 992 manual sets):** To address your concerns, we have carefully reviewed our initial submission and can confirm that the correct and consistently used number for both delineation sets is 993. We hope we have correctly interpreted your concern on this point.

3. **Regarding the origin of the 497 AI-Generated Sets:** Firstly, after 500 patients were enrolled, **3** images were dropped (Fig. 1). Subsequently, **one image from Center 05 failed to transfer to Center 01** (described in the caption of Fig. 1 and Supplementary Fig. 1). Therefore, for this specific image, manual, AI, and AI-assisted delineations were only generated at Center 05, and not at Center 01. This led to the final total number of manual **or** AI-assisted sets being **497 images×2 centers−1 image (in Center 01) =993 sets.**

To explain the source of these numbers more transparently, we have accordingly added this detailed information to the **Results section, Fig. 1, Supplementary Fig. 1 and their captions**.

---

15. Figure 1: While you explain the discrepancy in patient numbers in the figure caption, please also clarify this in the main text for consistency. Additionally, consider standardising the terminology, 'manual re-contouring' and 'manual delineation' refer to the same process but are currently used interchangeably. Consistent phrasing will improve clarity. Please clarify what you mean with "insufficient CT scan or model error" and also mention that this exclusion was performed in the main text.

Response：

We sincerely appreciate these kind suggestions. Accordingly, we have made the following revisions:

1.**Regarding the patient numbers and terminology:** To further clarify, we have now added a new subsection titled, **"Patient and Delineation Sets"** in the Results section. The discrepancy in patient numbers has now been explained both in this section and caption for Figure 1. Additionally, we have standardized the terminology to **"manual delineation"** in Figure 1 as you kindly suggested.

2.Regarding the dropout reasons:

In response to your valuable suggestion, we have clarified the details of this exclusion in the **"Patient and Delineation Sets"** subsection and the **caption for Figure 1:** "*Three of these patients were subsequently excluded from the final analysis due to technically inadequate data: two due to insufficient*

inferior scan coverage (omitting complete lungs and liver) and one due to a model generation failure."

Thank you again for your valuable suggestion, which has significantly improved the manuscript's readability and data transparency.

---

16. Figure 2b: The equation is written as M-Aa/M (i.e. M - (Aa/M)), but it seems you likely intended to use (M - Aa)/M. Please verify and correct the expression to ensure it reflects the intended calculation.

Response:

Thank you very much for your meticulous review. You are correct about the intended calculation, and we have now corrected the formula to **(M - Aa)/M in new Fig. 2c**.

---

17. Figure 2a, please add also the results after AI segmentation in this figure.

Response:

In response to your valuable suggestion, we have updated **Figure 2a–b** to include the AI segmentation results.

---

18. Figure 2c: please write CT-slices instead of CT-layers.

Response:

As you kindly suggested, we have corrected "CT layers" to "CT slices" in both the **new Fig. 2d** and its caption.

---

19. Figure 3a: please add more information on the Likert scale used in this assessment in the materials and methods section. Where the only possible choices: strongly satisfied, satisfied of neutral? Or was also a more negative choice possible? Was this assessment performed for every AI segmentation that was modified?

Response:

We sincerely appreciate your meticulous review and apologize for the lack of clarity.

1. Regarding the scale options: To address your question, our assessment did include negative choices ("Dissatisfied" and "Strongly Dissatisfied"); however, these options were not selected by any physician during the trial.

2. Regarding the assessment timing and scope: Prior to any manual revisions (i.e., AI-assisted delineation), physicians performed a single, overall assessment for each case, evaluating the entire set of 11 AI-generated OARs as a whole rather than rating each organ individually.

To provide this clarification, we have added the following description to the **"Evaluation Metrics" subsection** of the Methods: "*Prior to performing the AI-assisted task, the delineating physician rated their overall satisfaction with the initial set of 11 AI-generated OARs on a five-point Likert scale (from*

"Strongly satisfied" to "Strongly dissatisfied") for each case." **Figure 3a** and its caption have also been updated to reflect this.

---

20. How do you explain that the manual adjustment of an AI segmentation that was strongly satisfied/satisfied took more time than for an AI segmentation that was neutral (figure 3f)? I would expect that this would be the other way around.

Response：

Thank you for this insightful and thought-provoking question. This is indeed an interesting phenomenon and not an error in our data analysis.

First, as shown in the caption, the original Figure 4f (**now Figs. 4e–g**) presents a paired comparison (Strongly satisfied/satisfied vs. Neutral) where physicians evaluated and modified **the same initial AI segmentation.** We found that physicians who were satisfied with the same AI delineation achieved a significantly higher accuracy in their subsequent AI-assisted delineation (in terms of vDSC, HD95, sDSC etc.) than their "neutral" counterparts. **While the starting point was identical, reaching a more accurate endpoint required more detailed revisions and thus more time**, especially considering the greater effort needed for achieving near-perfect segmentation. Second, we found that for the same image, the "strongly satisfied/satisfied" physicians themselves tended to spend more time on manual delineation and also achieved higher accuracy (now Fig. 4g). This is consistent with the phenomenon observed in the AI-assisted delineation.

In summary, this may reflect **two distinct cognitive-behavioral patterns** in physician-AI interaction that could explain the origin of these different attitudes: Strongly satisfied/satisfied" physicians may apply stricter standards to their own contouring work (both manual and AI-assisted), but lower standards for the AI—while the "neutral" physicians do the opposite.

Inspired by your valuable question, we have incorporated a relevant discussion into **the fifth paragraph of the Discussion section**. We hope this explanation has adequately addressed your concerns.

---

21. In Figure 4b and 4f, the use of circles to indicate DSC improvement is visually unclear. Consider an alternative, more intuitive visual representation for this data.

Response：

Thank you for this valuable suggestion. To enhance the clarity of our presentation, we have replaced the previous dot plots with **heatmaps**. These new figures are based on paired analyses of the shared images between adjacent centers and illustrate the following key comparisons:

1. The effect of AI-assisted delineation versus manual delineation in reducing **inter-center** contouring variability (**new Figure 4c–e;** Supplementary Fig. 15).

2. The effect of AI-assisted delineation versus manual delineation in reducing **inter-physician** contouring variability (**new Figure 4f–i;** Supplementary Fig. 17).

We have also updated the relevant descriptions in the manuscript and believe that this revised format is more intuitive.

---

22. Figure 4b: why did you distinguish between previous and next centers? I would just present the differences in DSC between two centers independent of the order of segmentation.

Response：

Thank you for your constructive feedback. We agree that the "previous and next centers" terminology in the original figure was potentially confusing.

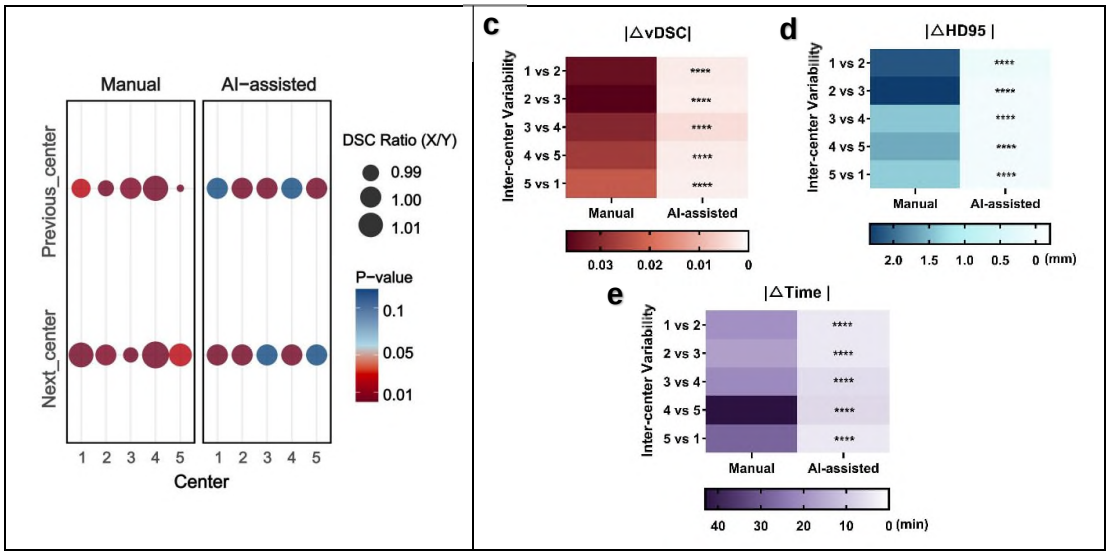
Our intention was to highlight that these were **paired comparisons** based on circulated CT scans **between adjacent centers**, as described in the "Study Design" section.

To address your suggestion for a more direct presentation, we have completely revised the figures. We replaced the dotplots with **heatmaps (new Figures 4c–e, and Supplementary Figs. 15)** that directly display the **absolute difference** in performance (for vDSC, HD95, sDSC etc.) between paired, adjacent centers. Additionally, the title has also been updated as "Inter-center Variability" to enhance clarity.

Based on this paired comparison, the revised visualization more intuitively and comprehensively demonstrates that AI-assisted delineation significantly **reduces inter-center differences** compared to manual delineation.

To visually demonstrate this improvement, we present a comparison of the original and revised figures below:

Original Figure 4b: Inter-institutional paired comparisons of DSC based on the same circulated images. DSC ratio (X/Y), the DSC of the center on the x-axis divided by that of the center on the y-axis.	**Revised Figures 4c–e:** Heatmaps showing the median inter-center variability, quantified as the absolute difference (|Δ|), in **(c)** vDSC, **(d)** HD95, and **(e)** contouring time between paired adjacent centers.



---

23. Figure 4f: please explain the abbreviations of the OARs in the caption.

Response：

Thank you for your kind reminder. Accordingly, we have now revised all figure captions to ensure that all abbreviations are clearly defined.

---

24. In Figure 5b: it only makes sense that OARs with a lower DSC have more DSC improvements, because their scores were lower to begin with. It would be more interesting to see DSC of the AI as a function of manual DSC and DSC of adjusted AI as a function of M-DSC to see whether more difficult structures are also more difficult for the network.

Response：

Thank you very much for this excellent insight. As you anticipated, we confirmed a significant **positive** correlation between AI/AI-assisted and manual delineation accuracy ($p<0.0001$; **Supplementary Fig. 18a–b**), indicating that the model and physicians potentially **face common challenges** when dealing with difficult-to-delineate OARs.

This new finding, together with our original observation that the greater AI improvements occurred where manual accuracy was lower (formerly Fig. 5b, now Supplementary Fig. 18c–d), provides a more fluid and logical transition to our subsequent exploration on the **advantages of AI over manual delineation in challenging segmentation scenarios,** as discussed in the section "Relationship and impact of AI assistance on manual delineation performance" (Supplementary Fig. 19a–e).

Thank you again for this valuable suggestion, which has helped us significantly improve the logical flow and the comprehensiveness of our analysis.

Supplementary data:

25. Table 1 and figure 2 of the supplementary data contain very relevant information and should preferably added to the main manuscript. Figure 2a of the main manuscript could be adjusted to add also the AI data. Than no additional figure is necessary. To see the manual, AI and AI-assisted results in one figure would be very insightful.

Response：

Thank you for your valuable suggestion. We have moved the original Supplementary Table 1 to the main manuscript **(now Table 1).** Furthermore, in the **revised Figure 2a–b and Supplementary Figs. 4 and 13**, we now provide a comprehensive, multi-metric comparison of the performance of manual, AI, and AI-assisted delineations.

---

26. The supplementary data contain many analyses and detailed tables, which are not all cited in the main manuscript. Now it is not easy for the reader to draw conclusions from all this information. Some adjustments are necessary:

Response：

Thank you for this very helpful feedback, as this was also a major concern for us. We initially provided detailed tables to simultaneously present our analysis results and ensure data transparency, but we agree that this approach compromised the readability of our results.

Following your suggestions, we have: (1) **converted the statistical tables in the Supplementary Materials into figures;** (2) enriched our analyses with **additional metrics** (e.g., HD95, sDSC); and (3) provided the raw data for all figures in the **Source Data** file to maintain transparency, in line with the journal's policy. We believe this new format significantly improves the clarity and accessibility of our findings.

---

a. present the most relevant information on contouring accuracy quantified in DSC or VRI in figures (like figure 2a in the main paper). Figures can be interpreted much easier than those busy tables. I suggest that you at least make figures for supplementary tables 3, 6, 15 and 21. Also graphs for the contouring times makes a fast interpretation easier than tables.

Response：

Thank you for your helpful guidance. In response, we have made the following revisions:

(1) We revised **Fig. 2a–b** and added **Supplementary Fig. 4 and 13** to demonstrate comparisons of manual, AI, and AI-assisted delineation using **multiple parameters** (vDSC, HD95, sDSC, VRI).

(2) We designed and created **new figures (e.g., Supplementary Figs. 2–20)** to replace the busy tables and demonstrate new analysis in the Supplementary Materials.

(3) We plotted figures for all time-related analyses (e.g., **Supplementary Figs. 12a–c, 15c and 16e**).

---

b. Add some text in the supplemental materials where you summarize the main findings and conclusions from the tables.

Response：

Thank you for this valuable suggestion. As mentioned in our previous response, we have replaced most supplementary tables with more intuitive figures (e.g., **Supplementary Figs. 2–20**). For the

few remaining tables (e.g., Supplementary Tables 1–2), we have incorporated summaries of their main findings into the **main text** to enhance readability. We hope that we have effectively addressed the concerns regarding the presentation of our data.

---

c. Consider removing some less relevant supplemental materials.

Response:

Thank you for your helpful suggestion to streamline our supplementary data. As we have now incorporated more robust metrics like **HD95 and sDSC** based on your guidance, we have **removed** the original tables concerning the Difference of Volume **(DV)**, Recall Rate **(RR)**, and Precision Rate **(PR)**. We believe that this refinement helps to focus reader's attention on the most robust and widely accepted metrics.

---

27. Add a table and graph for the comparison of DSC in CT-scans with and without contrast.

Response:

Thank you again for this valuable suggestion. We have accordingly added a new graph **(Supplementary Fig. 10)** to compare the delineation performance on contrast-enhanced versus non-contrast CTs across all three methods (manual, AI, and AI-assisted), using metrics including not only **vDSC** but also **HD95** and **sDSC**. In addition, following your prior guidance to reduce busy tables, the corresponding data table has been uploaded in the Source Data file.

---

28. In all tables concerning contouring time the unit of time is missing: please add this. I guess it should be minutes.

Response:

Thank you again for your meticulous review. We sincerely apologize for this oversight. As you correctly pointed out, the unit is **minutes (min).** We have now added the unit to the new figures (e.g., Supplementary Figs. 12, 15, and 16) that replaced the original time-related tables. Furthermore, we have double-checked all figures and tables throughout the manuscript to ensure all units are correctly specified.

---

29. Supplementary figure 3: it seems that the physician did not want to take enough time anymore to contour the lungs manually at the end of the study. Maybe because using the AI contours was much easier? Please add this in your discussion.

Response:

We sincerely appreciate this insightful suggestion. We strongly agree this is an important potential reason and have revised the manuscript accordingly. The **third-to-last paragraph of the Discussion section** now includes the following:

"*Conversely, for physicians with decreased manual accuracy, their delineation was often characterized by over- or under-segmentation at clear OAR boundaries and increased overlap among OARs (Supplementary Fig. 20). This decline may be attributed not only to fatigue from prolonged manual contouring, a limitation not experienced by AI,* ***but also*** *potentially to a*

behavioral shift; as physicians grew accustomed to the convenience of AI assistance, their motivation for time-consuming manual contouring may have diminished. Therefore, regularly reviewing contouring errors against atlases, ensuring adequate rest to avoid fatigue-induced errors, and mitigating the risks of over-reliance on AI tools remain essential in clinical physician-AI interaction."

---

30. Supplementary Method 1: data preprocessing: what was the resolution and slice thickness of the used CTs? What was the voxel size of the images after resampling?

Response：

We appreciate you pointing out the need for these details. Accordingly, we have added the following to **Supplementary Method 1:**

*"The native in-plane resolution (pixel spacing) ranged from **0.73 to 1.37 mm (median, 1.13 mm)**, and the slice thickness varied from **2.0 to 5.0 mm (median, 3 mm)** … Firstly, all CT slices were resampled to **1 × 1 mm** pixel spacing using nearest-neighbor interpolation to ensure spatial consistency across images."*

Furthermore, we have created **Supplementary Table 5** to provide detailed information about the imaging cohort for model development, such as in-plane resolution, number of slices, slice thickness, and contrast enhancement status.

---

31. Add more information on the deep learning autocontouring model used in this study in Method 2. What was the performance of the model in the test set? What kind of post-processing was used. Etc. Can the code of this model be shared?

Response：

Thank you for these important questions. Accordingly, we have addressed your points as follows:

1. We have substantially expanded the information on the model's development in **Supplementary Method 2**, in line with our response to your **Major Comment 2**.

2. The model's performance (vDSC and HD95) on the training, validation, and test sets, is now detailed in **Supplementary Table 6** and referenced in **Supplementary Method 2.**

3. We have added a new subsection titled "**Post-processing**" to Supplementary Method 2 to clarify these steps. The text reads:

"The segmentation outputs were post-processed to reduce false-positive results. The pipeline comprised three steps. Firstly, multiple consecutive morphological dilation and erosion operations were performed on the segmentation results of each organ to remove internal holes (less than 150 pixels). Next, the largest connected component of each organ was retained to isolate the primary anatomical structure. Finally, the segmentation maps were restored to their original size using the chest-region coordinates recorded during the preprocessing stage."

4. Regarding code availability, we fully agree that sharing code is important for scientific transparency and reproducibility. However, our model ("iCurveE") is a **commercial software product** (with Class III clinical approval from China's NMPA, No. 20253210066). Due to regulatory and intellectual property policies, we are **unable** to share the source code. We hope that

the substantial details provided in our replies to your **Major Comment 2** **and** **Detailed Comments 30–31** significantly enhance the model's transparency. We appreciate your understanding of this situation.

---

32. Supplementary table 33: what kind of AEs and SAEs during the CT imaging process are you referring to? Were the contours used in the clinic?

Response：

Thank you very much for this important question. First, we have revised the caption for the supplementary table (now **Supplementary Table 8**) to read:

"AEs were defined as unfavorable medical occurrences during CT imaging, such as a significant allergic reaction or severe anxiety, that led to scan interruption. SAEs were defined as events resulting in significant health deterioration or life-threatening injury. No such events were recorded during the study."

Second, this was a non-interventional clinical trial. As detailed in the **Study Design** section, the research protocol **did not mandate** the use of any trial-generated contours for clinical treatment, leaving this decision to the independent clinical judgment of the attending physician.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response：

We sincerely thank you for your detailed and insightful review. While the comments are not individually attributed, we wish to specifically thank you for your important contribution as a co-reviewer. Your suggestions have been invaluable in helping us improve the comprehensiveness of our trial's evaluation, the clarity of the data presentation, and the overall readability of our manuscript. We are very grateful for your efforts and believe that you will help many more studies and authors through this platform in the future.

Reviewer #3 (Remarks to the Author):

The study addresses a timely and clinically relevant topic, and I commend the authors for conducting a prospective, multicenter evaluation of an AI-based segmentation tool in routine radiotherapy workflows. However, the authors failed to leverage this opportunity to demonstrate the potential impact of AI in clinical practice.

Strengths:

1. Prospective multicenter design: The use of five clinical centers and a large, diverse group of physicians enhances the generalizability and clinical relevance of the findings.

2. Robust annotation and evaluation protocol: The study is meticulously designed, involving expert-defined ground truth annotations and comparative evaluation across manual, AI, and AI-assisted methods.

3. Comprehensive performance assessment: Multiple organs at risk (OARs) were assessed using key performance metrics (e.g., Dice similarity coefficient, time efficiency, and volumetric revision index), providing a thorough evaluation of the segmentation tool.

Response：

We sincerely thank you for your detailed review and critical suggestions. Following your valuable guidance, we have made extensive revisions to address the key issues you raised, supplementing the manuscript with the model's technical details, development information, benchmarking, dosimetric validation, and multi-parameter evaluation. We acknowledge that our initial submission, which focused primarily on reporting the results of this prospective multicenter clinical trial, was insufficient regarding these important points. We believe we have now substantially improved the comprehensiveness of our evaluation and the transparency of the details, and we thank you again for your invaluable guidance in strengthening our study.

Major Comments of Weaknesses

1. Description of the deep learning model

The segmentation model (iCurveE) appears to be a commercial tool, but the manuscript lacks detailed description of the underlying architecture, training datasets, or performance benchmarking against widely adopted open-source models (e.g., nnUNet). Inclusion of these details would allow readers to assess the novelty and generalizability of the model.

Response：

We sincerely thank the reviewer for this insightful comment. As you pointed out, iCurveE is indeed a commercial auto-segmentation tool, which has received **Class III clinical approval** (the highest risk category) from China's National Medical Products Administration (NMPA) (No. 20253210066). As our initial submission was centered on **reporting the outcomes of this prospective, multicenter clinical trial (NCT05787522)**, the technical details and performance benchmarks of the model itself were not sufficiently elaborated. In response, we have incorporated substantial new analyses and additional data into the manuscript. Specifically, we have:

1. Expanded the information of model architecture and development in **Supplementary Methods 1–2 and Supplementary Fig. 22.**
2. Detailed the imaging information of the model development cohort (n=827 scans) of iCurveE in **Supplementary Table 5.**
3. Reported the model's baseline performance on the training, validation, and test sets from the development cohort in **Supplementary Table 6.**
4. Benchmarked the iCurveE model against two open-source models: (a) an nnU-Net model trained from scratch using our development cohort and (b) the pre-trained TotalSegmentator. The results of these comparisons are presented in **Supplementary Figs. 2–3** and are now discussed in the new

section titled **“Performance Benchmark of the iCurveE Model”**.

Guided by your constructive feedback, we believe the manuscript now better elucidates the model's technical specifications and standalone performance.

---

2. Clinical relevance beyond segmentation accuracy

While the study convincingly demonstrates improvements in segmentation efficiency and accuracy, it does not evaluate the downstream impact on radiation treatment planning or patient outcomes. Metrics such as changes in dose-volume histogram parameters or plan quality would significantly enhance the clinical relevance of the findings.

Response：

We sincerely thank you for this important suggestion. Accordingly, we have now performed a dosimetric evaluation on 40 randomly selected patients. We calculated dosimetric deviations by comparing **38 clinically relevant DVH parameters** for each contouring method (manual, AI, and AI-assisted) against those derived from the ground-truth contours, using the fixed dose distribution from the clinical treatment plans. The methodology has been detailed in the “**Dosimetric Validation**” subsection of the Methods.

Our results show the mean dosimetric deviations were significantly lower for AI-assisted (4.2%) and AI-only (5.8%) delineations compared to manual delineation (11.1%). The AI and AI-assisted methods significantly reduced deviations in 13 (34.2%) and 23 (60.5%) of the DVH parameters, respectively **(Fig. 6a–b)**. Furthermore, we observed a significant correlation between higher geometric accuracy and smaller dosimetric deviations across all cancer types **(Fig. 6c–f)**. These results are now detailed in the Results subsection **“Reduced dosimetric deviations with AI-assisted delineation.”**

In line with your valuable suggestion, we have also emphasized the importance of evaluating patient outcomes in the **final paragraph of the Discussion**: *“However, rigorous validation through well-designed, prospective, interventional clinical trials with follow-up for patient outcomes, despite the formidable challenges, is crucial to ensure the safety and clinical applicability of future models.”*

Thank you again for this valuable guidance, which has markedly enhanced the comprehensiveness and clinical relevance of our study.

---

3. Justification for trial design

The rationale for using a cross-over trial design is not clearly articulated, especially given that patients were treated based on ground truth contours. Similarly, the purpose of cross-center contouring is unclear—whether it is intended to assess inter-physician variability, model generalizability, or end-user learning effects.

Response：

We sincerely thank you for this insightful question, which helps us better articulate the methodological design of our study.

1. Rationale for the cross-over design: This design was primarily to enable each physician to serve as their **own control** for a direct paired comparison between the two main clinical methods (manual vs. AI-assisted), while the randomized sequence minimized potential biases from contouring order.

2. Rationale for the cross-center contouring design: As correctly surmised, this design aimed to comprehensively assess the **inter-center and inter-physician variability** in manual and AI-assisted delineation. By having physicians from two different centers delineate **the same image sets**, the design enabled direct paired comparisons between: 1) physicians from adjacent centers (e.g., Fig. 4c–e); 2) experienced and junior physicians (e.g., Fig. 4f–i); and 3) physicians with differing satisfaction (e.g., Fig. 3e–g). This design effectively validated the value of iCurveE-based AI-assisted delineation **in reducing contouring variability**.

3. Rationale for the Ground Truth (GT): This study established a **multi-expert consensus GT** for all images to serve as a rigorous benchmark. In contrast, previous **retrospective studies** often use clinical treatment contour as the GT.[11-13] The generation of such a GT—manual delineation or revision of auto-segmentation by a single physician—is analogous to the "manual" and "AI-assisted" arms of our study and is also a source of heterogeneity. Furthermore, as an observational clinical trial (NCT05787522), our protocol **did not mandate** the clinical use of the time-consuming GT delineations. To ensure timely patient care, this decision was left to the discretion of the treating physician.

Following your valuable suggestions, we have added the corresponding clarifications to the **Study Design (Methods) and Results sections**. This has undoubtedly helped us enhance the clarity and depth of our methodological description.

---

4. Evaluation metrics

While Dice coefficient is widely used, it may not be sufficiently sensitive for large or irregularly shaped organs. Supplementing with additional metrics such as Hausdorff distance, mean surface distance, or dosimetric impact (e.g., changes in Dmax, Dmean) would provide a more comprehensive evaluation.

Response：

Thank you for this important guidance. We agree that supplementing our volumetric metrics with **distance- and surface-based analyses**, as also recommended by recent NRG Oncology guidance,[14] provides a more comprehensive evaluation.

Following your valuable advice, we have now calculated the 95th percentile Hausdorff Distance **(HD95)** and the surface Dice Similarity Coefficient **(sDSC)** for all 2,483 contour sets (manual, AI, and AI-assisted) on our multicenter image cohort (n=500). Furthermore, to assess the dosimetric impact of these geometric variations, we calculated **38 clinically relevant DVH parameters** for each delineation method (new Fig. 6), with the selection guided by the clinically widely used

QUANTEC guidelines and the Timmerman Table.[15,16]

We have now integrated these new results into the **main text (e.g., revised Figures 2-6) and Supplementary Materials (e.g., Supplementary Figs. 2–19)** to provide a more comprehensive assessment of contour accuracy.

Detailed Comments

Below are the review comments from Early Career Researcher who co-reviewed this paper:

The manuscript still needs a clearer justification for developing a 2-D Res-SE-Net ("iCurveE") rather than leveraging the many open-source or commercial solutions that already delineate thoracic OARs. To make the technical contribution unambiguous, please add an evaluation with three readily available comparators:

(Please incorporate these experiments or, if infeasible, provide a detailed rationale (with citations) for omitting them. Clarifying the performance gap between iCurveE and these baselines will substantially strengthen the paper's claim of technical and clinical novelty.)

1. Open-source baseline (trained from scratch). Train a standard nnU-Net on the identical multicentre cohort and report Dice, HD95 and inference time. This will show whether your network design itself, not merely the dataset, drives the reported gains.

Response：

Thank you for this further guidance. This is very helpful for clarifying the technical contribution of our model.

To address this point, we have performed an extensive comparative evaluation. We trained a standard nnU-Net model from scratch using the identical training cohort employed for iCurveE (Supplementary Table. 5) and its default, automated configuration;[17] and subsequently tested its performance on our multicenter study cohort.

This comparison revealed that iCurveE yielded significantly higher performance metrics for **6 OARs** (54.5%) (e.g., bilateral humeral heads and lungs, aorta, and liver), as measured by both vDSC and HD95. In contrast, **nnU-Net** performed better for the **trachea**. While the inference time for iCurveE was not prospectively collected in our trial, we have provided the inference time for the nnU-Net model (mean ± SD, 26.64 ± 23.60 s) in the Source Data file for your reference.

To present these findings, we have added a new section to the **Results ("Performance Benchmark of the iCurveE Model")** and included a detailed comparison in **Supplementary Fig. 2.**

---

2. Pre-trained, publicly released tool. Run Total Segmentator (which already contours all but the whole humerus vs. humeral head) on the test set. A quantitative comparison will reveal whether fine-tuning an existing model could achieve comparable accuracy and time-savings without developing a new architecture.

Response：

Thank you for the valuable suggestion to benchmark iCurveE against open-source models. As suggested, we evaluated the pre-trained **TotalSegmentator** on our multicenter cohort.

Assessed by both vDSC and HD95, iCurveE demonstrated superior accuracy in **5 of the 8 (62.5%)** OARs supported by both models ($p<0.0001$), while TotalSegmentator **does not support** the segmentation of the left and right humeral heads and the bronchial tree **(Supplementary Fig. 3).**

We attribute these advantages to our model's specialized training; both the development and multicenter study datasets consisted of planning CTs from thoracic cancer patients, with ground truth contours delineated by radiation oncology experts following radiotherapy guidelines. In contrast, the performance of TotalSegmentator may be suboptimal on scans with pathology or abnormal anatomy structures.[18]

Therefore, iCurveE demonstrates clear advantages in terms of both contouring accuracy and scope. These comparative results are now presented in **Supplementary Fig. 3** and in the **Results section (“Performance Benchmark of the iCurveE Model”).**

---

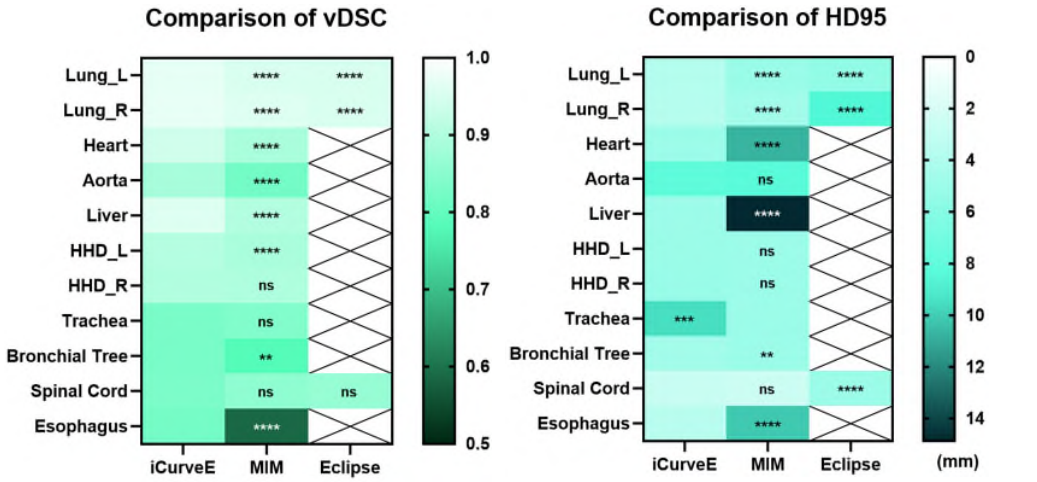
3. Commercial FDA-cleared TPS solutions. Where licences permit, include at least one widely used system that offers integrated OAR contouring such as MIM contour Protege, RayStation, or Eclipse. Such a benchmark would highlight your results in clinical application and underscore any added value for busy radiotherapy departments.

Response：

We thank the reviewer for this valuable suggestion to benchmark iCurveE against commercial TPS solutions. Although our institution is equipped with MIM (v7.3.3) and Eclipse (v15.5), we **do not have licenses** for their deep-learning based auto-contouring modules, such as MIM Contour Protégé or AI-Rad Companion Organs RT.

Nevertheless, we agree with your constructive opinion and conducted a relevant analysis. We imported the iCurveE training cohort to create a custom atlas for MIM's atlas-based segmentation. We then tested on **50** randomly selected cases from our multicenter cohort using the built-in segmentation tools of both systems: **MIM atlas-based segmentation** and **Eclipse model-based segmentation.** Our results showed that with the exception of the trachea, right humeral head, and spinal cord from MIM, **iCurveE outperformed** both commercial models in at least one key metric (vDSC or HD95) for all other OARs.

However, we are inclined **not to include** this analysis in the manuscript, primarily because these models **are not** the latest, deep-learning-based solutions, and also due to potential commercial sensitivities. We hope our rationale is understandable and that the two previous responses provide a sufficiently robust validation of iCurveE's technical contribution.



Performance benchmark of iCurveE against commercial TPS for OAR auto-segmentation. Heatmaps comparing the performance of iCurveE with MIM Atlas-based Segmentation and Eclipse model-based segmentation for 11 organs at risk (OARs). Performance was evaluated using (a) volumetric Dice similarity coefficient (vDSC) and (b) 95th percentile Hausdorff Distance (HD95). Asterisks denote statistically significant **inferior** performance in paired comparisons (i.e., lower vDSC or higher HD95), while a crossed-out box indicates the model's inability to segment the organ. ns, not significant; *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001. HHD_L, left humeral head; HHD_R, right humeral head.

---

4. Dosimetric Evaluation: Even though the authors explicitly acknowledge the lack of dosimetric comparison, While the geometric improvements reported for AI-assisted contours are highlighted, the manuscript stops short of demonstrating their clinical relevance. Since the proposed pipeline is intended for "treatment planning", please include a dosimetric comparison between plans generated with (i) ground-truth contours, (ii) manual physician contours, and (iii) AI/AI-assisted contours. At minimum, report standard dose-volume metrics, D_mean, D_max, D_min, D1, D50, and volume-based indices such as V1, V10, V20, V50, for the eleven OARs. A subset analysis on 20–30 representative cases would already clarify whether a 4–5 percentage-point Dice increase (e.g., 0.85 → 0.90) yields a tangible reduction in dose to critical structures.

Response：

Thank you for this constructive feedback. In response, we performed a dosimetric evaluation on 40 randomly selected patients. The methodology and corresponding results are detailed in the newly added subsections, **"Dosimetric Validation"** and **"Reduced dosimetric deviations with AI-assisted delineation",** respectively.

1. Regarding the dosimetric comparison:

Following your important guidance and established clinical guidelines,[15,16] we calculated dosimetric deviations by comparing **38 clinically relevant DVH parameters** for each contouring method (manual, AI, and AI-assisted) against those derived from the ground-truth contours. Our results showed that the mean dosimetric deviations were significantly smaller for AI-assisted (4.2%) and

AI-only (5.8%) delineations compared to manual delineation (11.1%). Furthermore, the AI-assisted method significantly reduced deviations in 23 (60.5%) of the DVH parameters **(Fig. 6a–b)**.

However, while this direct comparison is a commonly used method that reduces the influence of physicist-dependent plan optimization,[19-21] we fully agree with your guidance that generating distinct treatment plans for each contour set would offer a more comprehensive simulation of how delineation inaccuracies affect treatment plan dosimetry. Therefore, we have acknowledged and elaborated on this **limitation** in the Discussion section.

2. Regarding the dosimetry-accuracy sub-analysis:
We sincerely thank you for this insightful and clinically relevant suggestion. Considering that inaccurate OAR delineation can lead to either over- or underestimation of DVH parameters—and not just a simple increase[5]—we inferred that your original intention was for us to "clarify whether a Dice increase yields a tangible reduction in dose **deviation** to critical structures."

Accordingly, we performed this analysis and confirmed your suggestion, revealing a significant, albeit weak, correlation between higher geometric accuracy and smaller dosimetric deviations ($R^2 \approx 0.2$; $p<0.0001$) **(Fig. 6c–f)**. This weak correlation ($R^2 \approx 0.2$) is likely attributable to the fact that the dosimetric deviation depends more on the **distance** of the inaccuracy from the prescription dose line than on the overall contour accuracy. This finding is consistent with our previous dosimetric studies,[5] and we have now added a corresponding clarification in the **limitation section** of the Discussion.

We hope we have understood and addressed your suggestions appropriately. Thank you again for your invaluable guidance in enhancing the comprehensiveness and clinical relevance of our study.

---

5. Distance-based accuracy: The authors acknowledge that HD95 or surface distance were excluded because the decision pre-dated NRG Oncology guidance, leaving only volume metrics (DSC, VRI). For small, serial OARs (e.g. esophagus, brachial plexus) border errors can be clinically more important than volumetric overlap. A limited HD95 analysis on 20–30 test cases would strengthen the claim that AI assistance improves clinically important accuracy

Response：

We sincerely thank the reviewer for this critical feedback and apologize for the oversight.

To rectify this, we follow your valuable guidance and have now calculated the **HD95** and the **surface DSC (sDSC) for all 2,483 contour sets** (manual, AI, and AI-assisted) on our multicenter image cohort (n=500).

Furthermore, we have fully integrated these new results into the **main text (e.g., revised Figures 2-6) and Supplementary Materials (e.g., Supplementary Figs. 4–16)** to provide a more comprehensive assessment of contour accuracy.

---

6. Ablation Study for training cohort: The model was trained on 827 scans from only two Chinese hospitals, yet tested on five. An ablation in which those two centers are withheld from evaluation (true external test) would quantify domain-shift robustness.

Response：

Thank you for this important question. We would like to clarify that our study design inherently included a **true external test.** Specifically, the 827-scan model development cohort (from SYSUCC and SDCIH) **has no center, patient, or delineator overlap** with the 500-scan, five-center cohort of this study, as pointed out in the **Model information** subsection.

We recognize that the presence of "Sun Yat-sen University" in the names of both a training institution (SYSUCC) and a test institution (The Fifth Affiliated Hospital of Sun Yat-sen University (FHSU)) may have caused confusion. To enhance clarity, we have now **used unique abbreviations** for each hospital throughout the manuscript.

Furthermore, the standalone model performance evaluation across the five centers, presented in **Supplementary Table 1** and the revised **Figure 4**, serves as the quantification of the model's domain-shift robustness you requested. We believe this clarification fully addresses your concern.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response：

We would like to extend our special thanks for your professional and insightful feedback. Given the long-term enrollment and execution required for our prospective trial, the technical novelty of our model in such a rapidly advancing field could be a valid concern. Your invaluable guidance on conducting a benchmark comparison against state-of-the-art models was instrumental in addressing this. Furthermore, incorporating your suggestions on dosimetric validation and multi-parameter evaluation has significantly enhanced the comprehensiveness and robustness of our study. We sincerely appreciate your meticulous review and its crucial contribution to our work.

References

1. Li, G., Wu, X. & Ma, X. Artificial intelligence in radiotherapy. *Semin Cancer Biol* **86**, 160-171 (2022).
2. Busch, F., *et al.* Navigating the European Union Artificial Intelligence Act for Healthcare. *NPJ Digit Med* **7**, 210 (2024).
3. Lekadir, K., *et al.* FUTURE-AI: international consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ (Clinical research ed.)* **388**, e081554 (2025).
4. Pham, T. Ethical and legal considerations in healthcare AI: innovation and policy for safe and fair use. *R Soc Open Sci* **12**, 241873 (2025).
5. Niu, G.M., *et al.* Dosimetric analysis of brachial plexopathy after stereotactic body radiotherapy: Significance of organ delineation. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **190**, 110023 (2024).
6. 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. *International journal of radiation oncology, biology, physics* **81**, 1442-1457 (2011).
7. Jabbour, S.K., *et al.* Upper abdominal normal organ contouring guidelines and atlas: a Radiation Therapy Oncology Group consensus. *Practical radiation oncology* **4**, 82-89 (2014).
8. Weir, J. & Abrahams, P.H. *Weir & Abrahams' Imaging Atlas of Human Anatomy*, (Elsevier, 2020).
9. Huang, K., *et al.* Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. *Journal of applied clinical medical physics* **22**, 168-174 (2021).
10. Monnin, P., Sfameni, N., Gianoli, A. & Ding, S. Optimal slice thickness for object detection with longitudinal partial volume effects in computed tomography. *Journal of applied clinical medical physics* **18**, 251-259 (2017).
11. Ni, R., Han, K., Haibe-Kains, B. & Rink, A. Generalizability of deep learning in organ-at-risk segmentation: A transfer learning study in cervical brachytherapy. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **197**, 110332 (2024).
12. Podobnik, G., Ibragimov, B., Peterlin, P., Strojan, P. & Vrtovec, T. vOARiability: Interobserver and intermodality variability analysis in OAR contouring from head and neck CT and MR images. *Medical physics* (2024).
13. Wong, J., *et al.* Training and Validation of Deep Learning-Based Auto-Segmentation Models for Lung Stereotactic Ablative Radiotherapy Using Retrospective Radiotherapy Planning Contours. *Front Oncol* **11**, 626499 (2021).
14. 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. *International journal of radiation oncology, biology, physics* **119**, 261-280 (2024).
15. Bentzen, S.M., *et al.* Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues. *International journal of radiation oncology, biology, physics* **76**, S3-9 (2010).
16. Timmerman, R. A Story of Hypofractionation and the Table on the Wall. *International journal of radiation oncology, biology, physics* **112**, 4-21 (2022).
17. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J. & Maier-Hein, K.H. nnU-Net: a self-

configuring method for deep learning-based biomedical image segmentation. *Nat Methods* **18**, 203-211 (2021).

18. Wasserthal, J., *et al.* TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiol Artif Intell* **5**, e230024 (2023).
19. Ye, X., *et al.* Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. *Nature communications* **13**, 6137 (2022).
20. Hosny, A., *et al.* Clinical validation of deep learning algorithms for radiotherapy targeting of non-small-cell lung cancer: an observational study. *Lancet Digit Health* **4**, e657-e666 (2022).
21. Almberg, S.S., *et al.* Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology* **173**, 62-68 (2022).

Reviewer #1 (Remarks to the Author):

My compliments for your clear and extensive reply to our comments and for the efforts you have taken to improve the manuscript based. The quality of the manuscript has improved significantly and you have replied adequately to all our comments.

We appreciate your efforts in improving the visualisations for model comparisons. Especially supplental figure 3 with the boxplot representation gives a very clear overview of the differences between the different models. However, in Supplementary Figure 2, the chosen heatmap representation still makes it challenging to directly compare the iCurve model with nnU-Net and TotalSegmentator. We suggest you to combine Supplementary Figures 2 and 3 and show the results for both comparisons as boxplots.

Similarly, for Figures 4c–i, 6a–b, and Supplementary Figures 5, 6, 7, 8, 9, 10, 11, and 17, we suggest you to include the numerical values inside the boxes of the heatmaps to improve clarity and interpretability. Currently, while the heatmaps provide a clear visual overview, they do not fully convey the quantitative differences between models.

Response:

We are deeply grateful for your continued meticulous review and professional insights. We are delighted that our previous efforts have significantly improved the quality of our manuscript.

Following your further valuable suggestions, we have merged the original Supplementary Figures 2 and 3 into a **new Supplementary Figure 2**, utilizing boxplots to intuitively compare iCurveE with nnU-Net and TotalSegmentator.

Additionally, we have comprehensively embedded numerical values within the cells of **all heatmaps** in both the main text and Supplementary Materials to enhance the clarity and interpretability of the data presentation.

Thank you once again for your continued guidance. We believe the manuscript now fully meets the high standards of *Nature Communications*.

Reviewer #3 (Remarks to the Author):

The authors have sufficiently addressed my concerns, no more questions.

Response:

We sincerely appreciate your professional guidance and the positive assessment of our work. Your constructive feedback has been instrumental in refining the transparency, cutting-edge performance, and comprehensive evaluation of our model. Thank you once again!