Physics and Imaging in Radiation Oncology 34 (2025) 100738

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

External validation of an algorithm to detect vertebral level mislabeling and autocontouring errors

Tucker J. Netherton [b,*,1], Didier Duprez [a,1], Tina Patel [c], Gizem Cifter [b], Laurence E. Court [b], Christoph Trauernicht [a,2], Ajay Aggarwal [c,d,2]

[a] Division of Medical Physics, Tygerberg Hospital and Stellenbosch University, South Africa
[b] Department of Radiation Physics, Division of Radiation Oncology, University of Texas MD Anderson Cancer Center, United States
[c] Department of Radiotherapy Guy's & St Thomas NHS Foundation Trust London, the United Kingdom of Great Britain and Northern Ireland
[d] Department of Health Services Research and Policy, London School of Hygiene and Tropical Medicine, London, the United Kingdom of Great Britain and Northern Ireland

ARTICLE INFO

Keywords:
External validation of machine learning tools
Computed tomography
Vertebral level labeling
Image segmentation
Vertebral body segmentation

ABSTRACT

Background and Purpose: This work performs external validation of a previously developed vertebral body autocontouring tool and investigates a post-processing method to increase performance to clinically acceptable levels.
Materials and Methods: Vertebral bodies within CT scans from two separate institutions (40 from institution A and 41 from institution B) were automatically 1) localized and enumerated, 2) contoured, and 3) screened as a means of quality assurance (QA) for errors. Identification rate, contour acceptability rate, and QA accuracy were calculated to assess the tool's performance. These metrics were compared to those calculated on CTs from the model's original training dataset, and a post-processing technique was developed to increase the tool's accuracy.
Results: When testing the model without post-processing on external datasets A and B, accurate identification rates of 83 % and 92 % were achieved for vertebral bodies (C1-L5). Identification rate, contour acceptability rate and QA accuracy were reduced on both datasets compared to accuracies and rates measured on the model's orginal testing dataset. After algorithm adjustment, identification rate across all vertebrae increased on average by 4 % ($p < 0.01$) for dataset A and also 4 % on the dataset B ($p = 0.01$).
Conclusions: A post-processing adjustment within the machine learning pipeline increased performance of vertebral body localization accuracy to acceptable levels for clinical use. External validation of machine learning and deep learning tools is essential to perform before deployment to different insitutions.

1. Introduction

When cancer metastasizes to the vertebral bodies it can cause vertebral collapse, spinal cord compression, pain, and other chronic conditions which necessitate radiotherapy. The delivery of this radiotherapy is carefully planned by a multidisciplinary team often consisting of a dosimetrist, physicist, and radiation oncologist. Motivated by the time sensitive nature of such treatments, and the expertise needed to accurately label the spine, previous work has been performed to utilize machine learning techniques to autocontour vertebral bodies [1–3].

Computed tomography (CT) is the imaging modality most commonly used to identify vertebral bodies, and to use for 3-dimensional radiotherapy treatment planning. CT-based benchmark datasets for vertebral labeling and contouring are publicly available that include patients with normative anatomy and those with vertebral level variants [2]. For the purposes of this work, "labeling" refers to the correct classification of the vertebral body (e.g. C1, C2) with respect to its center-of-mass coordinates. Deep learning architectures using such data have gained high accuracy and have propelled the field of automated annotation and contouring forward. More recently, publicly available tools such as TotalSegmentor [4] can also contour individual vertebra bodies. However, to the best of our knowledge, such solutions do not verify (i.e.

* Corresponding author at: 1400 Pressler St, FCT8:6042, Houston, Texas 77030, United States.
E-mail address: tnetherton@mdanderson.org (T.J. Netherton).
[1] Authors Tucker Netherton and Dider Duprez are co-first authors.
[2] Authors Christoph Trauernicht and Ajay Aggarwal are co-senior authors.

https://doi.org/10.1016/j.phro.2025.100738
Received 8 July 2024; Received in revised form 12 February 2025; Accepted 24 February 2025
Available online 25 February 2025

perform secondary checks) vertebral body labeling nor contour quality. Such verification, or quality assurance (QA), is important, as wrong level labeling can lead to wrong level treatment [5,6].

Various machine learning based techniques for contour QA have been developed for the head and neck [7], pelvis [8], and spine [3]. These approaches mirror what is conceptually done during peer review and chart review, in which contour and plan quality is assessed [9,10]. Support vector machines, random forests, or other classification techniques can be used to flag errors made by autocontouring models caused by out-of-sample inputs. Automated QA for autocontouring can be used to notify the user when accuracy or performance of a tool is decreased. External validation of these QA approaches is essential before deployment of such tools to clinics with different patient populations and clinical practices [11].

Our previous work developed a robust method to label vertebral level in diagnostic and radiotherapy simulation CT images and was externally validated. Our subsequent work developed automated methods for contouring, radiotherapy planning, and an error detection (of labeled contours). However, while the automated contouring and error detection tools have been validated for clinical practice, they have not been validated on data from other clinics. Thus, the purpose of this work is to externally validate the autocontouring and quality assurance approaches on data from two different institutions. Motivation for this external validation comes from collaborative effort on the Radiation Planning Assistant (RPA), a web-based tool providing treatment planning to low resource clinics across the world [12,13]. In a survey of 15 clinics across 9 African countries, 50 % of those interested in using the RPA indicated that automated tools for palliative spine radiotherapy treatment planning would be useful [14]. This work evaluates the performance of the vertebral body autocontouring and QA tool on CT scans of patients from two different instiutions. The performance from each of these populations is compared to baseline results obtained from the institution where the original model was developed.

2. Materials and methods

2.1. Patient data

A total of 81 patients who received palliative radiotherapy to the spine were collected from Tygerberg Hospital, Stellenbosch University, Cape Town, South Africa (Institution A) (n = 41) and Guys and St Thomas Hospital, London, United Kingdom (Institution B) (n = 40) and under respective approved Institutional Review Board protocols. The criteria for inclusion were based on the recommendations of Netherton et al. [3], the inclusion criteria for the study required CT scans contain either the C1 or T1 vertebrae (so that manual vertebral body labelling could be verified). Institution A's CT scanner was a Phillips Big Bore; slice thickness was 3.0 mm for all scans and pixel spacing was 1.04 mm/pix on average (range = 0.99 to 1.37). Institution B's CT scanner was a Discovery CT590 RT; slice thickness was 2.5 mm for all patients and pixel spacing was 0.977 mm/pix for all patients.

2.2. Performance assessment

Contouring of all vertebral bodies is typically not performed in clinical practice due to the time-sensitive nature of palliative radiotherapy. Thus, for all 81 patient CT images, labeled contours (C1-L5) were automatically generated using the tool developed by Netherton et al. [3]. This tool first generates *two* sets of 3-dimensional coordinates of all vertebral bodies in the CT image using X-Net [1]. Like two experts with similar but different clinical training, these two models are separate deep learning "experts" trained on data from independent patient populations. Second, image patches are cropped around the centroids (from each of the primary and secondary models) and are passed into a deep learning-based contouring model. Third, image intensity features and pairwise contour similarity measures are calculated for each pair of corresponding vertebral body contours. Finally, these metrics are passed into a random forest binary classifier (the QA model) that is trained to predict if 1) contouring or vertebral body labeling errors exist or 2) if the contour quality and vertebral level is acceptable. The QA model operates under the assumption that if the two sets of contours are different, then there is likely an error present- this is thoroughly discussed and developed in prior work [3]. For reference, the performance values obtained by Netherton et al. for each step in this tool (the most updated version in clinical use) are listed in Table 1 [1,3].

Only the primary autocontours are used for clinical use, as the contours from the primary approach are slightly more accurate [1,3]. In this work, secondary contours are not seen by the user, but are only used as an input to the QA model so that a numerical score (from 0 to 1) can be obtained. The refinement of the QA model and quality of secondary contours made from the secondary models are thoroughly discussed in a prior work [3]. All primary autocontours generated by the approach described above were manually inspected by a clinical medical physicist in a treatment planning system (Eclipse, Siemens Medical Solutions, USA). Each primary vertebral body contour (n = 910 dataset A; 937 dataset B) was manually rated as acceptable or unacceptable, where acceptable meant that the contour, if used in autoplanning (e.g. autoconform MLC, jaws, isocenter) would result in clinically acceptable plan to treat the given vertebra. The performance of the quality assurance (QA) algorithm was scored as follows: true positive (TP) when the QA algorithm predicted a failure in labeling and/or contour quality and the medical physicist indicated that the vertebral contour was also unacceptable; true negative (TN) when the QA algorithm predicted no failure in labeling and/or contour quality and the medical physicist indicated that the vertebral contour was also acceptable; false positive (FP) when the QA algorithm predicted a failure in labeling and/or contour quality but the medical physicist indicated that the vertebral contour was acceptable; false negative (FN) when the QA algorithm did not predict a failure in labeling and/or contour quality but the medical physicist indicated that the vertebral contour was unacceptable.

2.3. Investigation of failures

If the accuracy of the QA model on each dataset was less than the performance achieved by Netherton et al. (Table 1, "baseline"), the

Table 1
Performance metrics.

	Data	IR	CR	QA A	QA R	QA S	QA F1
All regions	*Baseline*	*94 %*	*88 %*	*82 %*	*91 %*	*80 %*	*0.70*
	A	83 %	77 %	71 %	89 %	65 %	0.58
	$\overline{A}$	87 %	61 %	75 %	80 %	72 %	0.71
	B	92 %	84 %	76 %	87 %	74 %	0.54
	$\overline{B}$	96 %	87 %	75 %	74 %	75 %	0.43
Cervical	*Baseline*	*97 %*	*n/a*	*75 %*	*97 %*	*64 %*	*0.72*
	A	69 %	57 %	60 %	98 %	31 %	0.68
	$\overline{A}$	79 %	33 %	74 %	90 %	43 %	0.82
	B	92 %	74 %	74 %	99 %	66 %	0.66
	$\overline{B}$	94 %	75 %	68 %	87 %	62 %	0.58
Thoracic	*Baseline*	*95 %*	*n/a*	*78 %*	*93 %*	*72 %*	*0.70*
	A	87 %	83 %	72 %	79 %	71 %	0.49
	$\overline{A}$	89 %	68 %	74 %	73 %	75 %	0.64
	B	90 %	84 %	78 %	81 %	78 %	0.55
	$\overline{B}$	97 %	91 %	79 %	61 %	80 %	0.35
Lumbar	*Baseline*	*88 %*	*n/a*	*92 %*	*77 %*	*94 %*	*0.71*
	A	92 %	91 %	82 %	80 %	82 %	0.44
	$\overline{A}$	95 %	83 %	77 %	53 %	83 %	0.45
	B	97 %	96 %	73 %	38 %	74 %	0.10
	$\overline{B}$	97 %	96 %	75 %	38 %	77 %	0.11

cause of the performance decrease was investigated and a solution implemented to increase performance. The solution implemented was a post-processing step which "adapted" the algorithm for use on external data (Fig. 1). This adaptation preserved the predicted longitudinal coordinates from the localization model, but used the contour of the spinal canal (plus margin) to create an anterior bound for the thoracic coordinates predicted by the localization model. The anterior-posterior limit for the bounding region is 2.5 cm anterior to the midpoint of the spinal canal along the length of the thoracic spine. If any vertebral body coordinates are outside of this bounding region (primary or secondary coordinates), then each anterior-posterior coordinate is automatically adjusted to be placed on the anterior border of the bounding region (Fig. 2A). After application of the adaptation, the QA model was re-run to assess if the solution was effective and all metrics (listed in section 2.2) were re-calculated. Paired t-tests (one-sided, alpha = 0.05) were performed on quantitative metrics (before vs after the adaptation) to ascertain if the implemented adaptation techniques resulted in differences that were statistically significant. Due to the lack of ground truth data, no deep learning or machine learning models (described above) were re-trained.

2.4. Quantitative metrics

The quantitative metrics used in this work are identification rate, contour acceptability rate, QA accuracy, QA recall, QA specificity, and QA F1 score. These metric definitions are listed in Supplementary Material.

3. Results

3.1. Off-the-shelf performance

When the tool was ran on dataset A, IR of vertebral body regions was 69 % for cervical, 87 % for thoracic, and 92 % for lumbar regions. The QA algorithm accuracy was 60 % for cervical, 72 % for thoracic, and 82 % for lumbar regions. Compared to the performance metrics obtained on MDA data, IR was reduced by 11 % and QA accuracy was reduced by 14 % across all vertebral bodies when used on dataset A (Table 1).

When the tool was ran on dataset B, IR of vertebral body regions was 92 % for cervical, 90 % for thoracic, and 97 % for lumbar regions. The QA algorithm accuracy was 74 % for cervical, 78 % for thoracic, and 73 % for lumbar regions. Compared to the performance metrics obtained on baseline data, IR was reduced by 2 % and QA accuracy was reduced by 7 % across all vertebral bodies when used on dataset B (Table 1).

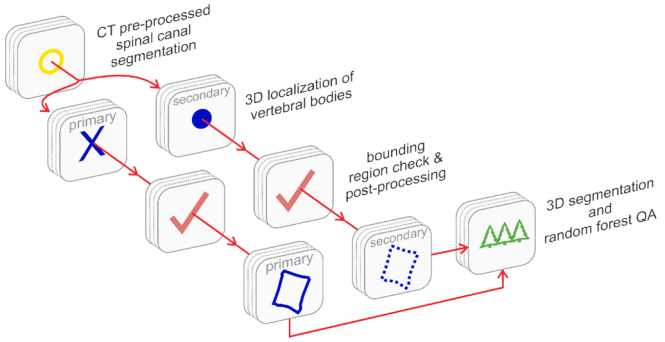

Fig. 1. Workflow diagram of the spine tool. After spinal canal segmentation, two sets of vertebral body coordinates are predicted. Then, the post-processing step modifies each set of coordinates if it fails the bounding region check. Primary and secondary autocontours are generated and passed to the QA model. Only the primary vertebral body autocontours, canal autocontour, and QA score are passed to the user.

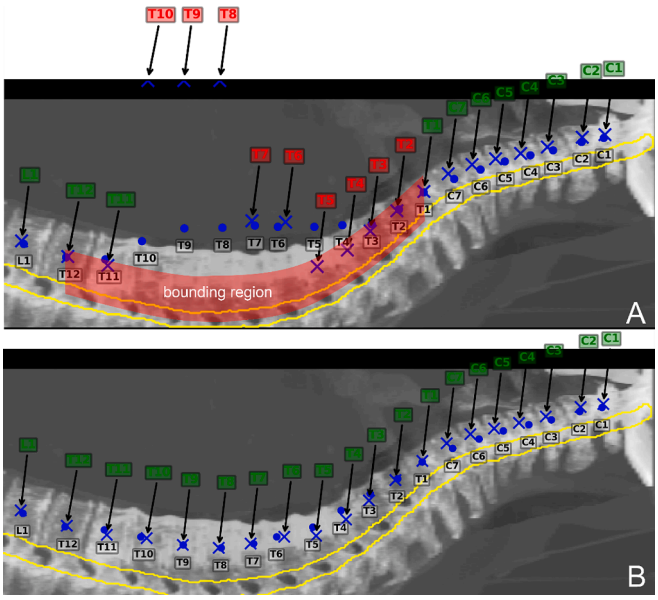

Fig. 2. Vertebral body localization coordinates with QA on a patient's scan from institution B. Blue X's and dots represent the coordinates predicted by the primary and secondary localization approaches, respectively. Green text indicates that the QA algorithm detected no labeling and/or contouring quality failure; red text indicates that the QA algorithm detected a failure in labeling and/or contouring quality. A is without the post-processing technique; B is with the post-processing localization technique. The red region in A from T1-T12 is the thoracic bounding region. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Post processing to improve localization of vertebral bodies

Decreases in vertebral body IR was visibly discernable on the thoracic region of many patients scans from institution B (Fig. 2A). This was not observed for the scans from dataset A. This technique improved IR on dataset A by 4 % ($p < 0.01$) on average and also 4 % on dataset B ($p = 0.01$) on average across all regions (Table 1). Interestingly, this technique improved average QA performance for dataset A only ($p = 0.02$). The impact of the technique with post-processing applied (aka "adapted technique") is listed for dataset A and dataset B patients in Table 1. Performance metrics for each individual vertebral level are listed in the Supplementary Material.

IR, mean identification rate; CR, contour acceptability rate; QA, quality assurance; A, accuracy; R, recall; S, sensitivity; F1, f1-score.

$\overline{A}$ and $\overline{B}$ indicate results after adaptation. *Baseline* indicates results from the unmodified models from Netherton et al. [1,3].

3.3. Adaptation assessment

The tool, with the adapted technique implemented, managed to accurately identify 79 % and 94 % (dataset A and dataset B) of cervical and 89 % and 97 % of thoracic and 95 % and 97 % lumbar spine vertebrae, even when portions of the scan did not include all seven cervical or all five lumbar spine vertebrae. The thoracic spine levels were more accurately labeled when utilizing the algorithm with the adaptation technique versus the unaltered algorithm and increased thoracic IR for dataset A by 2 % on average ($p = 0.05$) and by 7 % on average ($p = 0.02$) for dataset B. The tool, although providing high accuracy for majority of the patients, did struggle with atypical patients such as those with fused cervical spines, extra T13, variations in the number of lumbar spine levels. The tool also struggled to identify certain vertebral levels with several of the patients from both institutions due to extensive metastases throughout the spine (Fig. 3A) as well as two patients with very poor CT scan image quality. The extent of the metastases

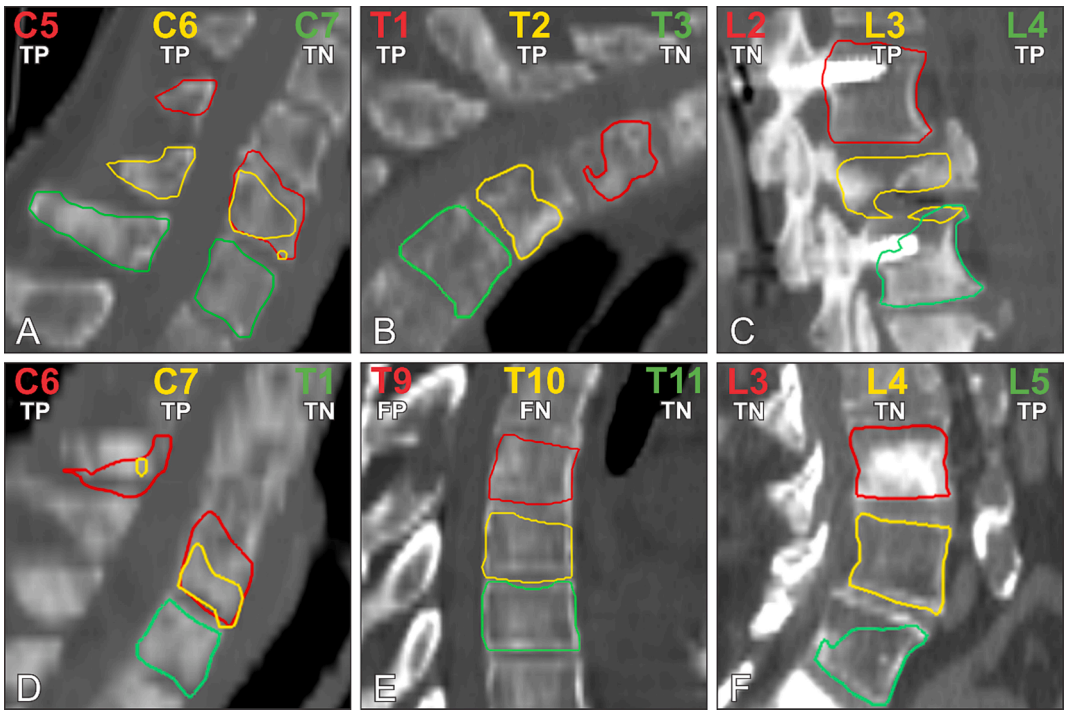

Fig. 3. Contours of vertebral bodies for dataset B (A,B,C) and dataset A (D,E,F) patients using the algorithm with the adaptation applied. TP, true positive; FP, false positive; TN, true negative; FN, false negative. Superior, middle, and inferior contours are in red, yellow, and green, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

and poor resolution made it extremely difficult, not just for the spinal tool, but also for the medical physicist, to correctly identify the appropriate level (Fig. 3D). The tool was able to perform surprisingly well despite these challenges and the QA algorithm was able to provide a handy guide or reassurance to the user when trying to identify the correct levels when presented with these complicated cases. In one such example the patient had both a collapsed vertebra and spinal hardware from vertebral stabilization (Fig. 3C). In this instance, the QA algorithm correctly identified the L3 and L4 contours had errors present. The QA tool was able to accurately flag vertebral levels that it determined to be either mislabeled or incorrectly contoured (e.g. Fig. 2B, T1). The QA tool was overall very conservative, with several cases involving vertebral levels that were flagged as unacceptable, however when checked by a professional it was found to be clinically acceptable. The QA tool managed to correctly flag problems with almost all the atypical patients.

4. Discussion

This work presented findings on the external validation of machine and deep learning algorithms used for spine autocontouring and QA. External validation of clinical, AI-based tools is crucial for patient safety, especially when such tools may be deployed where the burden of metastatic disease, image quality, and patient population is different to that of the population used to train the algorithms. A strength of this work is that rigorous external validation was performed using datasets from two continents. Furthermore, to our knowledge, there is no commercial system in existence which checks for quality or correct enumeration of vertebral body contours.

When the tool was used off-of-the shelf, performance metrics were decreased on average for all spinal regions for external data when compared to baseline data. All performance metrics, except for F1-score, were higher for dataset B than those from the dataset A when the tool was used off-the-shelf (Table 1). The metastatic burden in institution A scans was noticeably higher (through visual inspection) than those from institution B. Also, all slice thicknesses in dataset A (3 mm) were larger than the optimal slice thickness (2.5 mm) which was previously identified as the optimal value for this tool (to limit volume averaging) [3]. One important note is that the failure mode exhibited in Fig. 2A occurred in 11/40 patients in dataset B but only 1/41 patients in dataset A. After thorough investigation, it has remained elusive as to why the thoracic regions for 11 patients in dataset B had these distinct localization errors. However, the adaptation of the algorithm sufficiently corrected these failures and can be easily configured to be enabled or disabled (Fig. 1).The qualitative assessment of the algorithm performed by the medical physicist indicated that the adapted tool would be helpful for clinical practice and that review of the contours is recommended, as is performed in clinical practice.

The adaptation of the algorithm (performed via post processing of the localization coordinates) increased localization accuracy for both datasets by 4 %. However, QA performance values were increased for dataset B ($p = 0.03$), but decreased for dataset A ($p \ll 0.05$). For the random forest-based QA model to be effective, there must be a difference in autocontouring performance between primary and secondary methods. The random forest model was trained to identify when the primary autocontour does not match the secondary autocontour (see Netherton et al) [3]. Thus, an increase in IR, but decrease in QA accuracy (see Table 1, $\overline{B}$ cervical region and $\overline{A}$ lumbar region) may be attributed to primary and secondary autocontours that look more similar (from the applied adaptation technique), but are incorrect. While the localization and contouring models in this work were developed large amounts of data and deep learning techniques, the off-the-shelf performance on external datasets was decreased for both datasets. This was expected, as the training and testing data came from different patient populations. Overall, the ability of the random forest based QA model to detect failures (i.e. recall) was also decreased when tested on external datasets. This was also expected, as machine learning algorithms are known to have decreased performance on external datasets [15].

Due to the prevalence of some vertebral level variants, variability in body habitus, and presence of surgical implants (e.g. titanium hardware), it is not possible to due an extensive outlier analysis. However,

these datasets are representative of the average patient treated from each clinic for palliative radiotherapy and are suitable for use as benchmarks datasets for periodic quality assurance testing of the tool.

A limitation of this work, is that transfer learning or retraining of deep and machine learning models was not performed. Such approaches can be advantageous and would allow for model fine tuning and further refinement of specific performance metrics. This is a focus of future work and is associated with our development of new treatment approaches for the Radiation Planning Assistant, a tool that offers web-based treatment planning services at no cost to clinics with limited resources [13]. Future work will also include formal implementation mapping studies that investigate how new technology is best implemented and adapted for different patient populations and clinical practices.

In conclusion, although the deep learning and machine learning models evaluated in this work exhibit high accuracy, decreases in performance result when models are applied to external data. Adaptation of algorithms using post-processing increased performance of vertebral body localization accuracy to acceptable levels for clinical use. External validation of machine learning and deep learning tools is essential to perform before clinical AI algorithms are deployed.

CRediT authorship contribution statement

Tucker J. Netherton: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Didier Duprez:** Data curation, Formal analysis, Project administration, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tina Patel:** Data curation, Writing – original draft, Writing – review & editing. **Gizem Cifter:** Methodology, Writing – original draft, Writing – review & editing. **Laurence E. Court:** Conceptualization, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing. **Christoph Trauernicht:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing. **Ajay Aggarwal:** Conceptualization, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Laurence Court receives funding from Varian Medical Systems.

Acknowledgements

The authors would like to acknowledge the support of the Radiation Planning Assistant team, the Wellcome Trust, and the MD Anderson High Performance Computing team. Tucker Netherton would like to acknowledge the support of the National Institute of Health Loan Repayment Award.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2025.100738.

References

[1] Netherton TJ, Rhee DJ, Cardenas CE, Chung C, Klopp AH, Peterson CB, et al. Evaluation of a multiview architecture for automatic vertebral labeling of palliative radiotherapy simulation CT images. Med Phys 2020;47:5592–608. https://doi.org/10.1002/mp.14415.

[2] Sekuboyina A, Husseini ME, Bayat A, Löffler M, Liebl H, Li H, et al. VerSe: a Vertebrae labelling and segmentation benchmark for multi-detector CT images. Med Image Anal 2021;73:102166. https://doi.org/10.1016/j.media.2021.102166.

[3] Netherton TJ, Nguyen C, Cardenas CE, Chung C, Klopp AH, Colbert LE, et al. An automated treatment planning framework for spinal radiation therapy and vertebral-level second check. Int J Radiat Oncol Biol Phys 2022;114:516–28. https://doi.org/10.1016/j.ijrobp.2022.06.083.

[4] Wasserthal J, Breit H-C, Meyer MT, Pradella M, Hinck D, Sauter AW, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. Radiol Artif Intell 2023;5:e230024. https://doi.org/10.1148/ryai.230024.

[5] Shah M, Halalmeh DR, Sandio A, Tubbs RS, Moisi MD. Anatomical variations that can lead to spine surgery at the wrong level: part II thoracic spine. Cureus 2020. https://doi.org/10.7759/cureus.8684.

[6] Shah M, Halalmeh DR, Sandio A, Tubbs RS, Moisi MD. Anatomical variations that can lead to spine surgery at the wrong level: part III lumbosacral spine. Cureus 2020. https://doi.org/10.7759/cureus.9433.

[7] Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, et al. Automatic detection of contouring errors using convolutional neural networks. Med Phys 2019;46:5086–97. https://doi.org/10.1002/mp.13814.

[8] Rhee DJ, Akinfenwa CPA, Rigaud B, Jhingran A, Cardenas CE, Zhang L, et al. Automatic contouring QA method using a deep learning–based autocontouring system. J Appl Clin Med Phys 2022;23:e13647. https://doi.org/10.1002/acm2.13647.

[9] Cardenas CE, Mohamed ASR, Tao R, Wong AJR, Awan MJ, Kuruvila S, et al. Prospective Qualitative and quantitative analysis of real-time peer review quality assurance rounds incorporating direct physical examination for head and neck cancer radiation therapy. Int J Radiat Oncol 2017;98:532–40. https://doi.org/10.1016/j.ijrobp.2016.11.019.

[10] Ford E, Conroy L, Dong L, de Los Santos LF, Greener A, Gwe-Ya Kim G, et al. Strategies for effective physics plan and chart review in radiation therapy: report of AAPM Task Group 275. Med Phys 2020;47:e236–72. https://doi.org/10.1002/mp.14030.

[11] Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J 2021;14:49–58. https://doi.org/10.1093/ckj/sfaa188.

[12] Court LE, Aggarwal A, Jhingran A, Naidoo K, Netherton T, Olanrewaju A, et al. Artificial intelligence–based radiotherapy contouring and planning to improve global access to cancer care. JCO Glob Oncol 2024:e2300376. https://doi.org/10.1200/GO.23.00376.

[13] Court L, Aggarwal A, Burger H, Cardenas C, Chung C, Douglas R, et al. Addressing the global expertise gap in radiation oncology: the radiation planning assistant. JCO Glob Oncol 2023:e2200431. https://doi.org/10.1200/GO.22.00431.

[14] Dykstra MP, Netherton T, Mohammed BA, Lasebikan N, Kibudde S, Mallum A, et al. Treatment planning workflow and perceptions about automated contouring and treatment planning in Africa. Int J Radiat Oncol 2024;120:e748–9. https://doi.org/10.1016/j.ijrobp.2024.07.1644.

[15] Yu AC, Mohajer B, Eng J. External validation of deep learning algorithms for radiologic diagnosis: a systematic review. Radiol Artif Intell 2022;4:e210064. https://doi.org/10.1148/ryai.210064.