JAMA Network Open

Original Investigation | Health Informatics

Evaluation of Deep Learning to Augment Image-Guided Radiotherapy for Head and Neck and Prostate Cancers

Ozan Oktay, PhD; Jay Nanavati, MS; Anton Schwaighofer, PhD; David Carter, PhD; Melissa Bristow, MS; Ryutaro Tanno, PhD; Rajesh Jena, MD; Gill Barnett, MD; David Noble, MD; Yvonne Rimmer, MD; Ben Glocker, PhD; Kenton O'Hara, PhD; Christopher Bishop, PhD; Javier Alvarez-Valle, MS; Aditya Nori, PhD

Abstract

IMPORTANCE Personalized radiotherapy planning depends on high-quality delineation of target tumors and surrounding organs at risk (OARs). This process puts additional time burdens on oncologists and introduces variability among both experts and institutions.

OBJECTIVE To explore clinically acceptable autocontouring solutions that can be integrated into existing workflows and used in different domains of radiotherapy.

DESIGN, SETTING, AND PARTICIPANTS This quality improvement study used a multicenter imaging data set comprising 519 pelvic and 242 head and neck computed tomography (CT) scans from 8 distinct clinical sites and patients diagnosed either with prostate or head and neck cancer. The scans were acquired as part of treatment dose planning from patients who received intensity-modulated radiation therapy between October 2013 and February 2020. Fifteen different OARs were manually annotated by expert readers and radiation oncologists. The models were trained on a subset of the data set to automatically delineate OARs and evaluated on both internal and external data sets. Data analysis was conducted October 2019 to September 2020.

MAIN OUTCOMES AND MEASURES The autocontouring solution was evaluated on external data sets, and its accuracy was quantified with volumetric agreement and surface distance measures. Models were benchmarked against expert annotations in an interobserver variability (IOV) study. Clinical utility was evaluated by measuring time spent on manual corrections and annotations from scratch.

RESULTS A total of 519 participants' (519 [100%] men; 390 [75%] aged 62-75 years) pelvic CT images and 242 participants' (184 [76%] men; 194 [80%] aged 50-73 years) head and neck CT images were included. The models achieved levels of clinical accuracy within the bounds of expert IOV for 13 of 15 structures (eg, left femur, κ = 0.982; brainstem, κ = 0.806) and performed consistently well across both external and internal data sets (eg, mean [SD] Dice score for left femur, internal vs external data sets: 98.52% [0.50] vs 98.04% [1.02]; *P* = .04). The correction time of autogenerated contours on 10 head and neck and 10 prostate scans was measured as a mean of 4.98 (95% CI, 4.44-5.52) min/scan and 3.40 (95% CI, 1.60-5.20) min/scan, respectively, to ensure clinically accepted accuracy, whereas contouring from scratch on the same scans was observed to be 73.25 (95% CI, 68.68-77.82) min/scan and 86.75 (95% CI, 75.21-92.29) min/scan, respectively, accounting for a 93% reduction in time.

CONCLUSIONS AND RELEVANCE In this study, the models achieved levels of clinical accuracy within expert IOV while reducing manual contouring time and performing consistently well across previously unseen heterogeneous data sets. With the availability of open-source libraries and reliable

(continued)

Key Points

Question Can machine learning models achieve clinically acceptable accuracy in image segmentation tasks in radiotherapy planning and reduce overall contouring time?

Findings This quality improvement study was conducted on a set of 242 head and neck and 519 pelvic computed tomography scans acquired for radiotherapy planning at 8 distinct clinical sites with heterogeneous population groups and image acquisition settings. The proposed technology achieved levels of accuracy within interexpert variability; statistical agreement was observed for 13 of 15 structures while reducing the annotation time by a mean of 93% per scan.

Meaning The study findings highlight the opportunity for widespread adoption of autosegmentation models in radiotherapy workflows to reduce overall contouring and planning time.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Abstract (continued)

performance, this creates significant opportunities for the transformation of radiation treatment planning.

JAMA Network Open. 2020;3(11):e2027426. doi:10.1001/jamanetworkopen.2020.27426

Introduction

Each year, more than half a million patients are diagnosed with cancer and receive radiotherapy either alone or in combination with surgery.[1,2] Intensity-modulated radiation therapy has become a key component of contemporary cancer treatment because of reduced treatment-induced toxic effects, with 40% of successfully cured patients undergoing some form of radiotherapy.[3] Development of personalized radiation treatment plans that match a patient's unique anatomical configuration of tumor and organs at risk (OARs) is a multistep process starting with the acquisition of cross-sectional images and the segmentation of relevant anatomical volumes within the images through to dose calculation and subsequent delivery of radiation to the patient.

The segmentation of the images represents a significant rate-limiting factor within this treatment workflow. Currently, this task is performed manually by an oncologist using specially designed software to draw contours around the regions of interest. While the task demands considerable clinical judgement, it is also laborious and repetitive, with contoured volumes needing to be constructed slice by slice across entire cross-sectional volumes. Consequently, it is an extremely time-consuming process, often taking up to several hours per patient.[4] It can create delays in the workflow that may be detrimental to patient outcomes, but it also comes with an increasing financial burden to the hospital. As such, there is a significant motivation to provide automated or semiautomated support to reduce overall segmentation time for process.

In addition to long contouring times, there are challenges that derive from a dependency on computed tomography (CT) scans as primary reference images for tumor and healthy tissue anatomy. The inherent limitation of CT images in terms of image contrast on soft tissues makes segmentation challenging, and there remains uncertainty in the exact extent of tumor and normal tissues. This introduces a further key challenge for manual contouring; it is well documented that there is as a source of interoperator variability (IOV) in segmentation.[5-11] Such variability can affect subsequent dose calculations, with the potential for poorer patient outcomes.[12] Likewise, it presents a concern in the context of clinical trials carried out across multiple hospital sites. In addition to time savings, automating contouring would offer potential for greater standardization.

There has been significant investment to establish autosegmentation techniques that aim to reduce time and variability. Recent efforts are exploring machine learning (ML) methods for autosegmentation of CT scans in radiotherapy.[13-16] While they achieve reasonable accuracy within the same-site data sets on which they are trained and evaluated, model performance is often compromised when deployed across other hospital sites. Such approaches can be further limited in adaptability to different clinical domains of radiotherapy. Restricting these algorithms to a single bodily region or a single hospital site with specific acquisition protocols limits the value and applicability of these approaches in real-world clinical contexts. Furthermore, integration of such tools into existing hospital workflows is often not considered. To address these limitations, we present a generic segmentation solution for both prostate and head and neck cancer treatment planning and demonstrate how it can be integrated into existing workflows.

Methods

Ethical Review of Study

All data sets were licensed under an agreement with the clinical sites involved and received a favorable opinion from the research ethics committee from the East of England–Essex research ethics committee and the Health Research Authority. Under the agreements between the parties, the clinical sites agreed to obtain all consents, permissions, and approvals. This study followed the Standards for Quality Improvement Reporting Excellence (SQUIRE) reporting guideline.

The proposed segmentation method is based on a state-of-the-art convolutional neural network (CNN) model, and the same methodology is applied to both prostate and head and neck imaging data sets. It uses a variant of the 3-dimensional (3D) U-Net model[17] to generate contours of the OARs from raw 3D CT images (eAppendix 2 in the Supplement).

Segmentation Objectives

For prostate cancer, the model focuses contouring the following 6 structures: prostate gland, seminal vesicles, bladder, left and right femurs, and the rectum. For the purposes of radiotherapy planning in prostate cancer, radiation oncologists consider the prostate gland to be the target volume, while the remaining structures are delineated as OARs. In the case of head and neck cancer, we used a subset of OAR structures defined by a panel of radiation oncologists,[18] based on their relevance to routine head and neck radiotherapy (**Table 1**). The proposed model is trained to automatically delineate these 9 structures on a given head CT scan.

Image Data Sets and Manual Annotations

We aggregated 519 pelvic and 242 head and neck CT scans acquired at 8 different clinical sites from patients diagnosed either with prostate or head and neck cancer, as outlined in eAppendix 1 in the Supplement. The scans show variation across sites due to differences in scanner type and acquisition protocols. For experimental purposes, the images are grouped into 2 disjoint sets: main and external, as outlined in eAppendix 1 in the Supplement. The former is intended to be used for model training and testing purposes; the latter is an excluded data set composed of images from 3 randomly selected clinical sites and used to measure the model's generalization capability to unseen data sets. The main data set does not contain any images from these 3 excluded sites, thereby enabling a

Table 1. Autosegmentation Performance on 3 Head and Neck Data Sets

	Dice score, mean (SD)								
				Globe		Parotid		SMG	
Data set	Brainstem	Mandible	Spinal cord	Left	Right	Left	Right	Left	Right
IOV-10[a]									
Annotator 1	89.3 (4.2)	98.6 (1.0)	92.9 (1.5)	96.4 (0.9)	96.5 (1.1)	92.7 (3.5)	92.7 (3.5)	92.3 (3.4)	92.3 (2.6)
Annotator 2	91.8 (2.0)	98.5 (0.5)	91.8 (2.3)	95.6 (1.3)	96.7 (1.1)	91.1 (4.3)	91.2 (3.7)	91.3 (4.7)	91.3 (5.4)
Annotator 3	89.6 (2.7)	96.9 (1.0)	81.9 (7.3)	96.5 (0.8)	95.7 (1.0)	88.2 (3.8)	90.1 (2.8)	91.6 (2.8)	90.3 (8.0)
Ensemble	88.5 (2.0)	97.0 (1.0)	87.7 (3.6)	94.8 (1.0)	94.5 (1.9)	88.5 (2.3)	87.8 (4.1)	87.0 (2.9)	85.1 (5.3)
Agreement between annotators, κ	0.831	0.971	0.836	0.927	0.939	0.838	0.845	0.848	0.836
Agreement between annotators and model	0.806	0.966	0.844	0.917	0.931	0.852	0.825	0.803	0.794
Main data set, ensemble[b]	85.0 (3.7)	95.7 (2.3)	84.0 (3.8)	92.9 (1.6)	93.1 (1.5)	87.9 (3.8)	87.8 (4.3)	87.5 (2.3)	86.7 (3.5)
External data set, ensemble[c]	84.9 (6.8)	93.8 (2.5)	80.3 (7.7)	92.7 (3.6)	93.3 (1.4)	84.3 (4.6)	84.5 (4.3)	83.3 (9.1)	78.2 (21.1)
External data set,[c] Nikolov et al[15]	79.1 (9.6)	93.8 (1.6)	80.0 (7.8)	91.5 (2.1)	92.1 (1.9)	83.2 (5.4)	84.0 (3.7)	80.3 (7.8)	76.0 (16.5)
External data set, radiographer[c]	89.5 (2.2)	93.9 (2.3)	84.0 (4.8)	92.9 (1.9)	93.0 (1.7)	86.7 (3.5)	87.0 (3.1)	83.3 (19.7)	74.9 (30.2)

Abbreviations: IOV, interobserver variability; SMG, submandibular glands.

[a] IOV-10 data set included 10 images. In the IOV study, a subset of the main data set was annotated multiple times by 2 radiation oncologists and a trained reader. Later, the proposed model was compared against each human expert. The statistical agreement between annotators and model were measured with Fleiss κ values.

[b] Main data set included 20 images.

[c] External data set included 26 images. For the external data set, the reference ground truth contours were delineated by an expert head and neck oncologist, and IOV between clinical experts was measured by comparing the reference contours with those produced by an experienced radiographer.[15]

masked evaluation to be performed on the external data set. The images were manually annotated by 2 clinically trained expert readers (R.J. and G.B.) and a radiation oncologist masked to the others' annotations; as such, all structures in each image were manually contoured by 1 expert and later reviewed by a separate oncologist. For further details of the manual contouring process, see eAppendix 1 in the Supplement. The external head and neck data set was formed by using the head CT scans released by Nikolov et al,[15] which is an open-source data set[19] for benchmarking head and neck CT segmentation models and was acquired in The Cancer Imaging Archive Cetuximab[20] and The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma studies.

Evaluation Metrics

To evaluate model performance we used the Dice coefficient[21] as a similarity metric, which quantifies the correspondence between pairs of volumetric segmentations for the same structure. Perfectly overlapping structures result in a Dice score of 100.00%, while a Dice score of 0.00% corresponds to complete lack of overlap. In addition to this, we measured the overlap between pairs of contours using Hausdorff and mean surface-to-surface distance metrics (in mm). The metrics are visually presented and described further in eAppendix 3 in the Supplement.

Statistical Analysis

An ensemble of CNN models were trained with different training and validation set splits from main data set while leaving out a fixed disjoint testing set (see eAppendix 2 in the Supplement for details). The agreement between contours generated by the model and expert readers was measured statistically with the Cohen and Fleiss κ[22] for single and multiple annotators, respectively. For each structure, an agreement score was computed on foreground pixels defined by a binary mask. This is intended to avoid a possible bias due to a large number of background pixels. Similarly, Bland-Altman plots[23] were generated to visualize the level of agreement on a patient level (eAppendix 3 in the Supplement). The performance differences observed between the main and external sites was statistically tested with the Mann-Whitney test.[24] The same model training setup was also deployed on the main head CT data set to train a head and neck model that can delineate OARs in the context of head and neck radiotherapy(Table 1). **Figure 1** shows qualitative assessment of contours predicted with the proposed models. Additionally, to identify any gross contouring mistakes, the segmentations were also compared in terms of geometric surface distances.

In a second set of experiments to test the generalization to data sets from unseen clinical sites, the previously trained pelvic and head and neck CT models were tested on their corresponding external data set (external), which was comprised of images acquired at 3 particular clinical sites that were excluded from the training and validation data sets (main). With this experiment, the aim was to assess the generalization of the trained models to unseen CT acquisition protocols and patient groups.

All statistical analyses were conducted using Python version 3.7.3 (Python Software Foundation), with scikit-learn package version 0.21.1 for the Cohen-Fleiss κ and scipy package version 1.3.1 for the Mann-Whitney test. Statistical significance was set at $P < .01$ for null hypothesis testing and κ > 0.75 for the agreement analysis. All tests were 2-tailed.

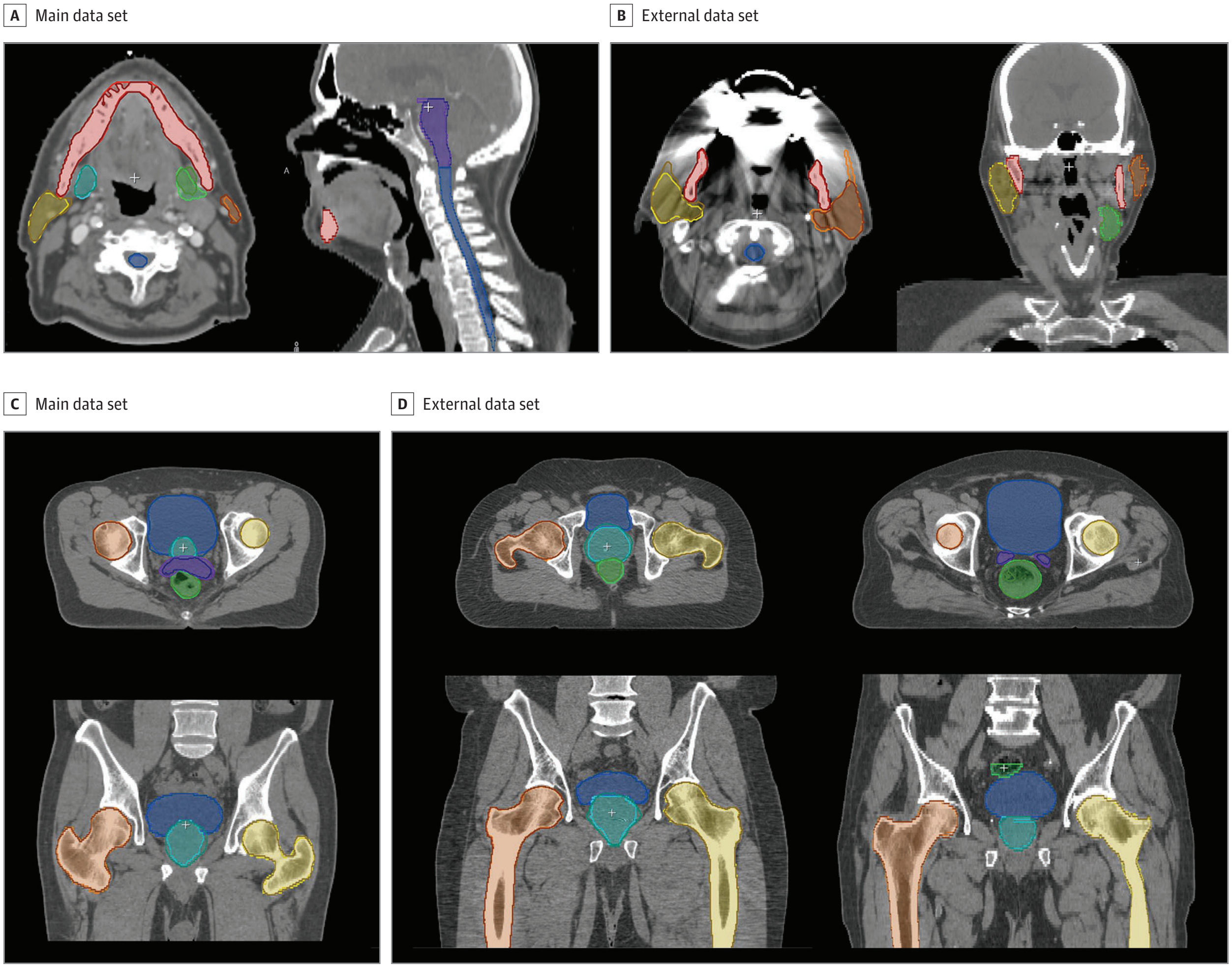
Results

A total of 519 participants' (519 [100%] men; 390 [75%] aged 62-75 years) pelvic CT images and 242 participants' (184 [76%] men; 194 [80%] aged 50-73 years) head and neck CT images were included. The prostate segmentation results (**Table 2**) show that the autogenerated organ delineations (ensemble) for prostate scans were consistent with the contours produced by clinical experts, with surface errors being within the acceptable error bound (eg, left femur, κ = 0.982). Results were consistent with head and neck segmentation results (eg, brainstem, κ = 0.806) (Table 1). Similarly, in validations on external data sets (Table 1 and Table 2), the model performed consistently well in both radiotherapy domains across multiple sites (eg, mean [SD] Dice score for left femur, internal vs external data sets: 98.52% [0.50] vs 98.04% [1.02]; *P* = .04), only with a slight performance drop on segmenting the submandibular glands due to low tissue contrast. Our observations of the segmentation errors tended to occur in the superior and inferior extent of tubular structures and in the interface between adjacent organs. However, we have not observed any inconsistencies that, if not corrected, could lead to significant errors in a treatment plan, as evidenced by the surface distance results. This is because the proposed postprocessing method does not allow inconsistencies at a distance from the anatomical structure by design.

IOV Analysis

An acceptable measure of performance is expected to be within the bounds of IOV found in human experts.[5,7] The IOV Dice scores and surface distances between 3 experts contouring 10 test images for each radiotherapy domain are provided in Table 1 and Table 2. For 14 of 15 structures, statistical agreement (ie, κ > 0.75) was observed between autogenerated contours and expert annotations. The reference contours were determined by applying a majority voting scheme using all 3 annotators. At least 2 experts must have agreed to imply that a structure is in fact present. For all the structures except SMGs, the similarity scores with ground truth achieved the criteria of being on-par with levels of expert IOV in contouring, as indicated by the κ values and Bland-Altman plots (eAppendix 3 in the Supplement) collected for the agreement analysis. Here we can see that for more

Figure 1. Qualitative Evaluation of Expert and Autogenerated Contours on Head and Neck Computed Tomography Scans



clearly defined structures with high contrast, such as the bladder and femurs, there is reasonably high consistency across the experts (κ > 0.96). But for lower contrast and deformable features, such as the prostate gland, seminal vesicles, and SMGs, we see a higher rate of variability because the organ boundaries are typically unclear in the presence of such adverse conditions (**Figure 2** and Table 2). A similar pattern of performance difference is seen on the contours generated by the model, where the same test images are segmented and compared qualitatively with the same reference contours (Figure 2).

Table 2. Autosegmentation Performance on 3 Pelvic Data Sets

	Dice score, Mean (SD)					
	Femur					
Data set	Left	Right	Bladder	Rectum	Prostate	SV
IOV-10[a]						
Annotator 1	98.79 (0.33)	98.72 (0.40)	97.31 (1.54)	90.42 (5.75)	89.83 (4.82)	83.47 (7.65)
Annotator 2	99.63 (0.12)	99.63 (0.11)	98.20 (0.65)	95.49 (1.90)	88.66 (6.67)	82.98 (11.71)
Annotator 3	99.51 (0.17)	99.43 (0.17)	98.10 (0.71)	91.78 (4.73)	85.44 (8.26)	78.02 (13.55)
Ensemble	98.94 (0.34)	98.92 (0.34)	97.00 (1.27)	89.90 (4.13)	88.05 (1.43)	81.18 (5.66)
Agreement between annotators, κ	0.985	0.984	0.962	0.864	0.787	0.685
Agreement between annotators and model, κ	0.982	0.981	0.959	0.852	0.820	0.732
Main data set, ensemble[b]	98.52 (0.50)	98.50 (0.58)	95.68 (2.56)	87.73 (4.03)	87.17 (3.70)	80.69 (5.91)
External data set, ensemble[c]	98.04 (1.02)	98.02 (1.13)	95.84 (1.82)	87.03 (3.01)	86.51 (4.74)	80.13 (7.00)
P value, main vs external data set[d]	.04	.04	.10	.07	.42	.91
Main data set, ensemble, MD	0.25 (0.09)	0.25 (0.10)	0.69 (0.20)	1.71 (0.86)	1.62 (0.52)	1.07 (0.41)
External data set, ensemble, MD	0.30 (0.16)	0.30 (0.18)	0.81 (0.37)	2.19 (1.19)	1.73 (0.58)	1.19 (0.56)
IOV-10 data set, ensemble, MD	0.15 (0.04)	0.15 (0.05)	0.56 (0.20)	1.48 (0.80)	1.43 (0.39)	0.96 (0.36)
IOV-10 data set, annotators, MD	0.10 (0.07)	0.11 (0.07)	0.40 (0.19)	1.03 (1.01)	1.41 (0.91)	1.07 (0.88)
Main data set, ensemble, HD	1.20 (0.22)	1.19 (0.25)	2.42 (0.67)	7.57 (5.54)	4.32 (1.77)	3.71 (1.47)
External data set, ensemble, HD	1.40 (0.56)	1.39 (0.71)	2.86 (0.78)	8.96 (6.71)	5.06 (2.09)	4.29 (2.35)
IOV-10 data set, ensemble, HD	1.03 (0.12)	1.02 (0.12)	2.85 (0.62)	6.64 (3.89)	4.07 (1.04)	3.64 (1.59)
IOV-10 data set, annotators, HD	0.74 (0.46)	0.72 (0.49)	2.45 (0.95)	6.30 (5.16)	5.27 (2.74)	5.24 (3.36)

Abbreviations: HD, Hausdorff distance; IOV, interobserver variability; MD, mean surface distance; SV, seminal vesicles.

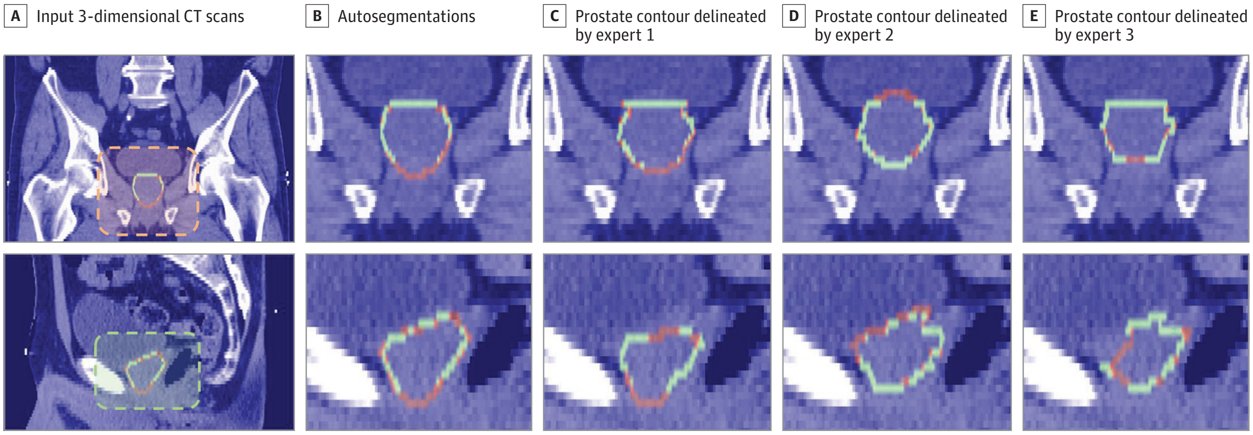
[a] IOV-10 data set included 10 images. In the IOV study, the results for each annotator are reported separately to show the distribution and how they compared with autogenerated contours. The statistical agreement between annotators and model were measured with Fleiss κ values.

[b] Main data set included 49 images.

[c] External data set included 83 images.

[d] The statistical significance of differences between Dice scores on external and main data sets was assessed with the Mann-Whitney test.

Figure 2. Interexpert Variability In Prostate Contour Annotations



A Input 3-dimensional CT scans

B Autosegmentations

C Prostate contour delineated by expert 1

D Prostate contour delineated by expert 2

E Prostate contour delineated by expert 3

CT indicates computed tomography.

Annotation Time for Treatment Planning

The clinical benefit of the models was assessed by comparing times to correct autogenerated contours with times to manually contour images from scratch. For this analysis, the head and neck IOV-10 and prostate IOV-10 data sets were used, in which the scans had varying imaging quality, ranging from good (15) to poor (5). An in-house annotation tool is used for both contouring and correction tasks. The tool features assistive contouring and interactive contour refinement modules that ease the contouring task while ensuring high segmentation accuracy. Manual segmentation of the head and neck scans for the same 9 OARs took a mean of 86.75 (95% CI, 75.25-98.29) min/scan for an expert reader and 73.25 (95% CI, 68.68-77.82) min/scan for a radiation oncologist. For the same scans, the review and correction time of autocontours was measured as 4.98 (95% CI, 4.44-5.52) min/scan for head and neck scans and 3.40 (95% CI, 1.60-5.20) min/scan for prostate scans, which are inspected and updated (if necessary) by the oncologist to ensure clinical accuracy required for treatment planning. This represented a mean 93% reduction in time. Among all 20 scans, the slowest correction time per scan was measured as 7.05 minutes because of low imaging quality. A mean inference time of 23 (95% CI, 20-26) seconds was taken to segment target and all OAR foreground pixels in full input CT scan.

Discussion

Several frameworks have been proposed for autosegmentation of head and neck [15,25] and pelvic organs.[13,16,26] In 2 studies,[13,16] the authors describe an approach for prostate and OAR segmentation, where organ localization is performed prior to segmentation. Their algorithm was validated on a data set of 88 CT scans. A similar cascaded autosegmentation approach was proposed in Wang et al[26] to delineate OARs in head and neck CT scans; this study was conducted by training (33 scans) and evaluating (15 scans) using the public data set released in an autosegmentation challenge.[27]

There have been efforts to show the potential clinical use cases of ML solutions for automatic OAR contouring. In contrast to previous work, in which evaluations were performed on small sets of homogenous images, we evaluated how ML solutions could lead to generalized performance across (1) different radiotherapy domains and (2) data sets from multiple sites. We aimed to demonstrate the robustness and generalizability of these solutions. More importantly, we found that integrating these models into clinical workflows could reduce the time required to prepare dose plans for treatment.

The models demonstrated performance generalizability across diverse acquisition settings while achieving good levels of agreement with expert contours. This could facilitate easier deployment in new clinical sites. Of further importance for any practical adoption of this technology across large scale health care systems is the ability to work across diverse clinical domains. We have shown how our approach, without any substantial changes, can enable the training of models in diverse radiotherapy domains, as demonstrated through applications in prostate and head and neck cancer. This is especially significant given the distinct imaging challenges associated with these different domains.

Practical adoption in clinical contexts is enhanced by incorporating the presented models into the existing workflow of radiation oncologists (**Figure 3**). The illustrated system has been implemented and evaluated by clinical experts working at Cambridge University Hospitals. In this workflow, CT scans are acquired from patients as they attend preparations for radiotherapy treatment. These scans are initially stored at the hospital's image database and later securely transferred via the gateway to the autosegmentation platform in the cloud after anonymizing them. Once the segmentation process is completed, resultant files are uploaded back to the hospital's image database, creating a seamless clinical workflow in which clinicians can review and refine contours in their existing contouring and planning tools.

Bringing these ML tools to the point where they can be meaningfully adopted in clinical practice requires a level of clinical accuracy commensurate with expert observers. While the models have performed well in this regard, in instances where the model performed poorly, the opportunity to manually correct the segmentations remains a necessary component of the presented workflow. The presented workflow enables oncologists to use their existing clinical systems for review and editing, which makes this technology more accessible across clinics because the existing workflows are maintained. At the same time, clinicians can inspect and edit contours in minutes rather than hours. Such time savings are significant even when considered only in absolute terms.

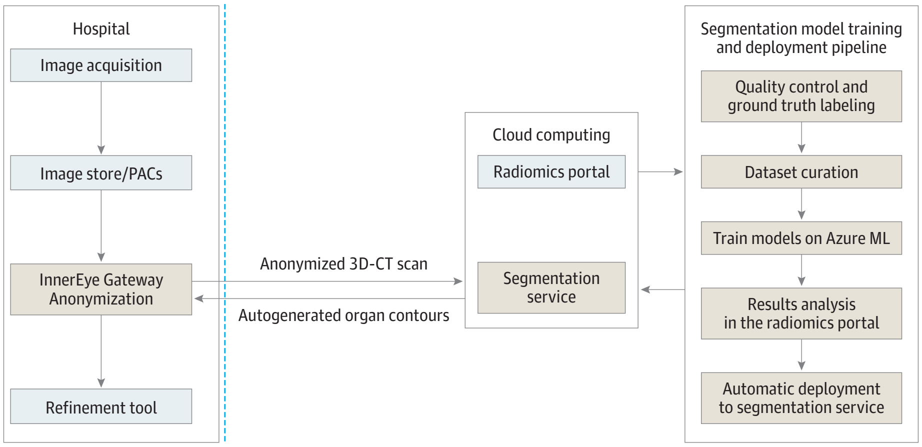
The source code used in this study is made publicly available.[28] This creates an opportunity for oncology centers to use this technology to train and deploy new models using their own data sets. In this way, users can include other normal tissue structures in the autocontouring pipeline, including cochlear and oral-cavity structures in head and neck cancer treatments. The availability of new public data sets and sharing across clinics is an important milestone in improving the performance of models and making them accessible. Similarly, image quality (IQ) assurance[29] is essential for reliable use of models. IQ assessment should be performed prior to model deployment[30] both at acquisition and processing time to filter out images with metal artifacts. Training models on a diverse set of data sets, as performed in this study, is an effective way to cope with low-contrast (eg, cone-beam CT) and high-noise images. External data set validation is also essential to measure such impacts; for instance, the images from the external head and neck data set used in this study contained severe beam-hardening artifacts.

More adaptive forms of radiotherapy, in which anatomy is resegmented and the dose plan reoptimized for each fraction, are regarded as a more ideal way to deliver treatment,[31] which has been challenging to adopt due to its heavy resource demand.[32] In that regard, the presented technology can enable continuous resegmentation and adaptive reoptimization of therapy to be adopted at scale. For instance, in the cases of hypofractionated regimens or emergency treatments, extension of these models to resegment anatomy on scans would have significant clinical utility to save time and allow patients to progress to treatment more quickly. Integration with technologies such as The Magnetic Resonance Linear Accelerator,[33] used for simultaneous imaging and dose delivery, could also potentially offer more adaptive forms of treatment to pinpoint the location of tumors at the time of treatment.

Limitations

This study has limitations. The data sets used in the IOV and annotation time experiments are smaller than the remaining evaluations presented in this study. For further statistical significance, these experiments shall be repeated with larger data sets with varying imaging quality. Additionally, surface and Dice metrics used in model evaluation do not always correlate with time savings in manual

Figure 3. Integration of the Proposed Segmentation Models Into Radiotherapy Planning Workflow


Hospital
Image acquisition
Image store/PACs
InnerEye Gateway Anonymization
Refinement tool
Anonymized 3D-CT scan
Autogenerated organ contours
Cloud computing
Radiomics portal
Segmentation service
Segmentation model training and deployment pipeline
Quality control and ground truth labeling
Dataset curation
Train models on Azure ML
Results analysis in the radiomics portal
Automatic deployment to segmentation service

contouring process.[15,34] This necessitates the design of new metrics that quantify segmentation errors by taking into account the cost of required user interaction to correct them.

Conclusions

This study found that ML-based autosegmentation reduces contouring time while yielding clinically valid structural contours on heterogeneous data sets for both prostate and head and neck radiotherapy planning. This is evidenced in evaluations on external data sets and IOV experiments conducted on a multisite data set. Overall, the approach contributes to the practical challenges of scalable adoption across health care systems through off-the-shelf extensibility across hospital sites and applicability across multiple cancer domains. Future ML studies validating the applicability of the proposed technology on other radiotherapy domains and larger data sets will be valuable for wider adoption of ML solutions in health care systems.

ARTICLE INFORMATION

Accepted for Publication: October 1, 2020.

Published: November 30, 2020. doi:10.1001/jamanetworkopen.2020.27426

Corresponding Author: Ozan Oktay, PhD, Health Intelligence, Microsoft Research, 21 Station Rd, Cambridge, CB1 2FB United Kingdom (ozan.oktay@microsoft.com).

Author Affiliations: Health Intelligence, Microsoft Research, Cambridge, United Kingdom (Oktay, Nanavati, Schwaighofer, Carter, Bristow, Tanno, Jena, Barnett, Glocker, O'Hara, Bishop, Alvarez-Valle, Nori); Department of Oncology, Cambridge University Hospitals NHS Foundation Trust, United Kingdom (Noble, Rimmer); now with Edinburgh Cancer Centre, Western General Hospital, Edinburgh, United Kingdom (Noble).

Author Contributions: Dr Oktay had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Concept and design: Oktay, Nanavati, Schwaighofer, Tanno, Jena, Noble, Glocker, Bishop, Alvarez-Valle, Nori.

Acquisition, analysis, or interpretation of data: Oktay, Nanavati, Schwaighofer, Carter, Bristow, Jena, Barnett, Noble, Rimmer, Glocker, O'Hara, Alvarez-Valle.

Drafting of the manuscript: Oktay, Nanavati, Schwaighofer, Bristow, Jena, Glocker, O'Hara, Alvarez-Valle, Nori.

Critical revision of the manuscript for important intellectual content: Oktay, Schwaighofer, Carter, Tanno, Jena, Barnett, Noble, Rimmer, Glocker, O'Hara, Bishop, Alvarez-Valle, Nori.

Statistical analysis: Oktay, Nanavati, Schwaighofer, Carter.

Obtained funding: Nori.

Administrative, technical, or material support: Oktay, Schwaighofer, Bristow, Jena, Noble, Bishop, Alvarez-Valle, Nori.

Supervision: Oktay, Schwaighofer, Jena, Barnett, Rimmer, Glocker, Alvarez-Valle, Nori.

Conflict of Interest Disclosures: Dr Jena reported receiving personal fees from Microsoft during the conduct of the study. Dr Noble reported receiving grants from Cancer Research UK and personal fees from Microsoft Research, Cambridge, during the conduct of the study. No other disclosures were reported.

Funding/Support: The research work reported in the manuscript was self-funded by Microsoft Research Cambridge.

Role of the Funder/Sponsor: The funder had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Pan HY, Haffty BG, Falit BP, et al. Supply and demand for radiation oncology in the United States: updated projections for 2015 to 2025. *Int J Radiat Oncol Biol Phys*. 2016;96(3):493-500. doi:10.1016/j.ijrobp.2016.02.064

2. Sklan A, Collingridge D. Treating head and neck cancer: for better or for worse? *Lancet Oncol*. 2017;18(5):570-571. doi:10.1016/S1470-2045(17)30269-3

3. Barnett GC, West CM, Dunning AM, et al. Normal tissue reactions to radiotherapy: towards tailoring treatment dose by genotype. *Nat Rev Cancer*. 2009;9(2):134-142. doi:10.1038/nrc2587

4. Vorwerk H, Zink K, Schiller R, et al. Protection of quality and innovation in radiation oncology: the prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study). *Strahlenther Onkol*. 2014;190(5):433-443. doi:10.1007/s00066-014-0634-0

5. Cazzaniga LF, Marinoni MA, Bossi A, et al. Interphysician variability in defining the planning target volume in the irradiation of prostate and seminal vesicles. *Radiother Oncol*. 1998;47(3):293-296. doi:10.1016/S0167-8140(98)00028-0

6. Cooper JS, Mukherji SK, Toledano AY, et al. An evaluation of the variability of tumor-shape definition derived by experienced observers from CT images of supraglottic carcinomas (ACRIN protocol 6658). *Int J Radiat Oncol Biol Phys*. 2007;67(4):972-975. doi:10.1016/j.ijrobp.2006.10.029

7. Dubois DF, Prestidge BR, Hotchkiss LA, Prete JJ, Bice WS Jr. Intraobserver and interobserver variability of MR imaging- and CT-derived prostate volumes after transperineal interstitial permanent prostate brachytherapy. *Radiology*. 1998;207(3):785-789. doi:10.1148/radiology.207.3.9609905

8. Fiorino C, Reni M, Bolognesi A, Cattaneo GM, Calandrino R. Intra- and inter-observer variability in contouring prostate and seminal vesicles: implications for conformal treatment planning. *Radiother Oncol*. 1998;47(3):285-292. doi:10.1016/S0167-8140(98)00021-8

9. Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. *Strahlenther Onkol*. 2012;188(2):160-167. doi:10.1007/s00066-011-0027-6

10. Gardner SJ, Wen N, Kim J, et al. Contouring variability of human- and deformable-generated contours in radiotherapy for prostate cancer. *Phys Med Biol*. 2015;60(11):4429-4447. doi:10.1088/0031-9155/60/11/4429

11. Valicenti RK, Sweet JW, Hauck WW, et al. Variation of clinical target volume definition in three-dimensional conformal radiation therapy for prostate cancer. *Int J Radiat Oncol Biol Phys*. 1999;44(4):931-935. doi:10.1016/s0360-3016(99)00090-5

12. Ohri N, Shen X, Dicker AP, Doyle LA, Harrison AS, Showalter TN. Radiotherapy protocol deviations and clinical outcomes: a meta-analysis of cooperative group clinical trials. *J Natl Cancer Inst*. 2013;105(6):387-393. doi:10.1093/jnci/djt001

13. Balagopal A, Kazemifar S, Nguyen D, et al. Fully automated organ segmentation in male pelvic CT images. *Phys Med Biol*. 2018;63(24):245015. doi:10.1088/1361-6560/aaf11c

14. Lou B, Doken S, Zhuang T, et al. An image-based deep learning framework for individualising radiotherapy dose. *Lancet Digit Health*. 2019;1(3):e136-e147. doi:10.1016/S2589-7500(19)30058-5

15. Nikolov S, Blackwell S, Zverovitch A, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. arXiv. Preprint published online September 12, 2018. Accessed October 29, 2020. https://arxiv.org/pdf/ 1809.04430.pdf

16. Wang S, He K, Nie D, Zhou S, Gao Y, Shen D. CT male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. *Med Image Anal*. 2019;54:168-178. doi:10.1016/j.media.2019.03.003

17. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells W, Frangi A, eds. *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Lecture Notes in Computer Science, vol 9351*. Springer; 2015. doi:10.1007/978-3-319-24574-4_28

18. Brouwer CL, Steenbakkers RJ, Bourhis J, et al. CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines. *Radiother Oncol*. 2015;117(1):83-90. doi:10.1016/j.radonc.2015.07.041

19. TCIA CT scan dataset. Accessed October 29, 2020. https://github.com/deepmind/tcia-ct-scan-dataset

20. Cancer Imaging Archive. Head-neck cetuximab. Published June 3, 2020. Accessed October 29, 2020. https://wiki.cancerimagingarchive.net/display/Public/Head-Neck+Cetuximab

21. Dice LR. Measures of the amount of ecologic association between species. In: *Ecology*. 1945;26(3):297–302. doi:10.2307/1932409

22. Lin L, Hedayat AS, Wu W. *Statistical Tools for Measuring Agreement*. Springer Science & Business Media, 2012."https://doi.org/10.1007/978-1-4614-0562-7"

23. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;1(8476):307-310. doi:10.1016/S0140-6736(86)90837-8

24. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Ann Math Stat*. 1947;18(1):50-60. doi:10.1214/aoms/1177730491

25. Zhu W, Huang Y, Zeng L, et al. AnatomyNet: deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy. *Med Phys*. 2019;46(2):576-589. doi:10.1002/mp.13300

26. Yueyue Wang et al Organ at risk segmentation in head and neck CT images using a two-stage segmentation framework based on 3D U-net. *IEEE Access*. 2019;7:144591–144602. doi:10.1109/ACCESS.2019.2944958

27. Raudaschl PF, Zaffino P, Sharp GC, et al. Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015. *Med Phys*. 2017;44(5):2020-2036. doi:10.1002/mp.12197

28. InnerEye Deep Learning. Accessed October 29, 2020. https://github.com/microsoft/InnerEye-DeepLearning/

29. Barrett JF, Keat N. Artifacts in CT: recognition and avoidance. *Radiographics*. 2004;24(6):1679-1691. doi:10.1148/rg.246045065

30. Tarroni G, Oktay O, Bai W, et al. Learning-based quality control for cardiac MR IMAGES. *IEEE Trans Med Imaging*. 2019;38(5):1127-1138. doi:10.1109/TMI.2018.2878509

31. Sonke JJ, Aznar M, Rasch C. Adaptive radiotherapy for anatomical changes. *Semin Radiat Oncol*. 2019;29(3):245-257. doi:10.1016/j.semradonc.2019.02.007

32. Heukelom J, Fuller CD. Head and neck cancer adaptive radiation therapy (ART): conceptual considerations for the informed clinician. *Semin Radiat Oncol*. 2019;29(3):258-273. doi:10.1016/j.semradonc.2019.02.008

33. Lagendijk JJW, Raaymakers BW, van Vulpen M. The magnetic resonance imaging-linac system. *Semin Radiat Oncol*. 2014;24(3):207-209. doi:10.1016/j.semradonc.2014.02.009

34. Valenzuela W, Ferguson SJ, Ignasiak D, et al. FISICO: fast image segmentation correction. *PLoS One*. 2016;11(5):e0156035. doi:10.1371/journal.pone.0156035

35. Acosta, O., Dowling, J., Drean, G. et al Multi-atlas-based segmentation of pelvic structures from CT scans for planning in prostate cancer radiotherapy. In: El-Baz AS, Saba L, Suri JS, eds. *Abdomen and Thoracic Imaging*. Springer, 2014:623–656. doi:10.1007/978-1-4614-8498-1_24

SUPPLEMENT.

eAppendix 1. Author Contributions and Data Set
eAppendix 2. Supplementary Methods
eAppendix 3. Supplementary Material
eReferences.