Large-scale crowdsourced radiotherapy segmentations across a variety of cancer anatomic sites: Interobserver expert/non-expert and multi-observer composite tumor and normal tissue delineation annotation from a prospective educational challenge

Kareem A. Wahid[a]*, Diana Lin[b]*, Benjamin E. Nelms[c], Renjie He[a], Mohammed A. Naser[a], Simon Duke[d], Michael V. Sherer[e], John P. Christodouleas[f,g], Abdallah S. R. Mohamed[a], Michael Cislo[b], James D. Murphy[e], Clifton D. Fuller[a]**, Erin F. Gillespie[b]**

[a]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
[b]Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY.
[c]Canis Lupus, LLC, Merrimac, WI, USA.
[d]Department of Radiation Oncology, Cambridge University Hospitals, Cambridge, UK.
[e]Department of Radiation Medicine and Applied Sciences, University of California San Diego, La Jolla, CA, USA.
[f]Department of Radiation Oncology, The University of Pennsylvania Cancer Center, Philadelphia, PA, USA.
[g]Elekta, Atlanta, GA, USA.

* co-first authors
** co-corresponding authors

Corresponding authors:

Clifton D. Fuller, MD, PhD
Department of Radiation Oncology
The University of Texas MD Anderson Cancer Center
1515 Holcombe Blvd.
Houston, TX, 77030
Email: cdfuller@mdanderson.org

Erin F. Gillespie, MD
Department of Radiation Oncology
Memorial Sloan Kettering Cancer Center
1275 York Ave, Box 22
New York, NY 10065
Email: gillespe@mskcc.org

Funding Statement: This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) through a Cancer Center Support Grant (CCSG; P30CA016672-44; P30CA008748). D.L. is supported by the Radiological Society of North America (RSNA) Research Medical Student Grant (RMS2116). K.A.W. is supported by the Dr. John J. Kopchick Fellowship through The University of Texas MD Anderson UTHealth Graduate School of Biomedical Sciences, the American Legion Auxiliary Fellowship in Cancer Research, and an NIH/National Institute for Dental and Craniofacial Research (NIDCR) F31 fellowship (1 F31DE031502-01). E.F.G. and J.D.M. received funding from the Agency for

Health Research and Quality (AHRQ R18HS026881). C.D.F. received funding from the NIH/NIDCR (1R01DE025248-01/R56DE025248); an NIH/NIDCR Academic-Industrial Partnership Award (R01DE028290); the National Science Foundation (NSF), Division of Mathematical Sciences, Joint NIH/NSF Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant (NSF 1557679); the NIH Big Data to Knowledge (BD2K) Program of the NCI Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825); the NCI Early Phase Clinical Trials in Imaging and Image-Guided Interventions Program (1R01CA218148); an NIH/NCI Pilot Research Program Award from the UT MD Anderson CCSG Radiation Oncology and Cancer Imaging Program (P30CA016672); an NIH/NCI Head and Neck Specialized Programs of Research Excellence (SPORE) Developmental Research Program Award (P50CA097007); and the National Institute of Biomedical Imaging and Bioengineering (NIBIB) Research Education Program (R25EB025787).

Conflict of Interest: D.L. is in a research fellowship funded by grants for research and education related to eContour.org. B.E.N. is the founder of ProKnow and advisor to Elekta AB. J.P.C. is an employee of Elekta AB. C.D.F. has received direct industry grant support, speaking honoraria, and travel funding from Elekta AB. E.F.G. is a co-founder of the educational website eContour.org. The other authors have no conflicts of interest to disclose.

Abstract

Clinician generated segmentation of tumors and healthy tissue regions of interest (ROIs) on medical images is crucial for radiotherapy workflows. However, interobserver segmentation variability has long been considered a significant detriment to the implementation of high-quality and consistent radiotherapy dose delivery. This has prompted the increasing development and utilization of computational approaches for automated ROI segmentation, in particular deep learning based methods. However, extant public segmentation datasets typically only provide segmentations generated by a limited number of annotators with varying, and often unspecified, levels of expertise. In this data descriptor, a large number of clinician annotators manually generated segmentations for numerous ROIs on computed tomography images across a variety of cancer sites (breast, sarcoma, head and neck, gynecologic, gastrointestinal) for the Contouring Collaborative for Consensus in Radiation Oncology challenge. Over 200 clinician annotators (verified experts and non-experts) contributed to the generation of these segmentation data using a standardized annotation platform (ProKnow). Subsequently, we converted image and segmentation data into Neuroimaging Informatics Technology Initiative format with standardized nomenclature for ease of use by the research community. In addition, we generated consensus segmentations for experts and non-experts using the simultaneous truth and performance level estimation method. These standardized, structured, and easily accessible data are a valuable resource for systematically studying interobserver variability in ROI segmentation applications with an unprecedented large number of clinician annotators.

Background & Summary

Since the advent of contemporary radiation delivery techniques for cancer treatment, clinician generated segmentation (also termed contouring or delineation) of target structures (e.g., primary tumors and metastatic lymph nodes) and organs at risk (e.g., healthy tissues whose irradiation could lead to damage and/or side effects) on medical images has become a necessity in the radiotherapy workflow [1]. These segmentations are typically provided by trained medical professionals, such as radiation oncologists. While segmentations can be performed on any imaging modality that provides sufficient discriminative capabilities to visualize regions of interest (ROIs), the current radiotherapy workflow prioritizes the use of computed tomography (CT) for ROI segmentation due to its ubiquitous nature and use in radiotherapy dose calculations. Subsequently, clinicians spend a large fraction of their time and effort generating ROI segmentations on CT imaging necessary for the radiotherapy workflow.

Interobserver and intraobserver variability are well-documented byproducts of the use of manual human-generated segmentations [2,3]. While consensus radiotherapy guidelines to ensure ROI segmentation quality have been developed and shown to reduce variability [4], these guidelines are not necessarily followed by all practicing clinicians. Therefore, segmentation variability remains a significant concern in maintaining radiotherapy plan quality and consistency. Recent computational improvements in machine learning, particularly deep learning, have prompted the increasing development and deployment of accurate ROI auto-segmentation algorithms to reduce radiotherapy segmentation variability [5–7]. However, for auto-segmentation algorithms to be clinically useful, their input data (training data) should reflect high-quality “gold-standard” annotations. While research has been performed on the impact of interobserver variability and segmentation quality for auto-segmentation training [8–11], it remains unclear how “gold-standard” segmentations should be defined and generated. One common approach, consensus segmentation generation, seeks to crowdsource multiple segmentations from different annotators to generate a high-quality ground-truth segmentation. While multi-observer public medical imaging segmentation datasets exist [12–17], there remains a lack of datasets with a large number of annotators for radiotherapy applications.

The Contouring Collaborative for Consensus in Radiation Oncology (C3RO) challenge was developed to engage radiation oncologists across various expertise levels in cloud-based ROI crowdsourced segmentation [18]. Through this collaboration, a large number of clinicians generated ROI segmentations using CT images from 5 unique radiotherapy cases: breast, sarcoma, head and neck, gynecologic, and gastrointestinal. In this data descriptor, we present the curation and processing of the data from the C3RO challenge. The primary contribution of this dataset is unprecedented large-scale multi-annotator individual and consensus segmentations of various ROIs crucial for radiotherapy planning in an easily accessible and standardized imaging format. These data can be leveraged for exploratory analysis of segmentation quality across a large number of annotators, consensus segmentation experiments, and machine learning model building. An overview of this data descriptor is shown in **Fig. 1**.

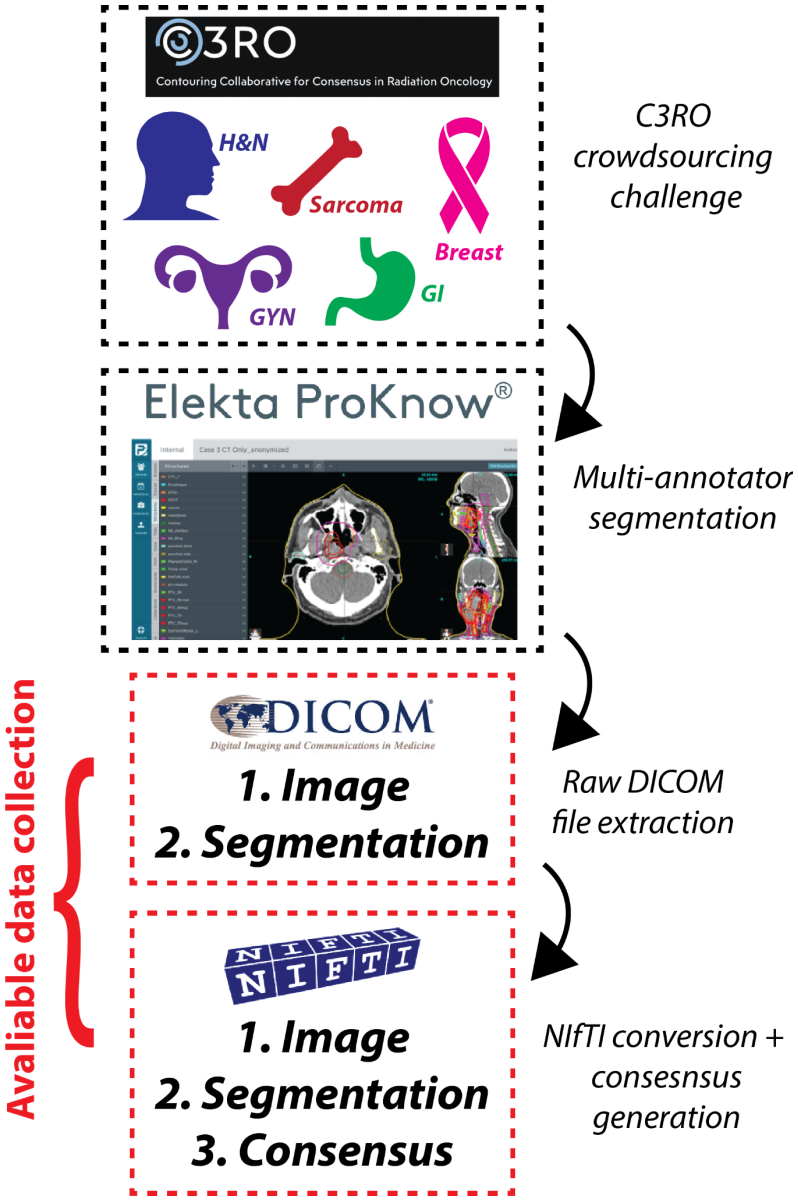

Figure 1. Data descriptor overview. Multi-annotator segmentations were generated for the Contouring Collaborative for Consensus in Radiation Oncology (C3RO) challenge. Imaging and segmentation data were extracted from the ProKnow cloud-based platform in Digital Imaging and Communications in Medicine (DICOM) format, which were then converted to Neuroimaging Informatics Technology Initiative (NIfTI) format for ease of use. Consensus segmentation data were then generated from the available multi-annotator segmentations. The provided data collection contains all original DICOM files along with converted NIfTI files.

Methods

Patient population. 5 separate patients who had undergone radiotherapy were retrospectively collected from our collaborators at various institutions. All patient data used was anonymized before transmission for use by the C3RO team, and therefore informed consent was not

required. Each patient had received a pathologically confirmed diagnosis of cancer of one of the following sites: breast (post-mastectomy intraductal carcinoma), sarcoma (malignant peripheral nerve sheath tumor of the left thigh), head and neck (oropharynx with nodal spread, [H&N]), gynecologic (cervical cancer, [GYN]), and gastrointestinal (anal cancer, [GI]). Clinical characteristics of these patients are shown in **Table 1**.

Table 1. Clinical characteristics for cases included in this data descriptor. Cases included breast, sarcoma, head and neck (H&N), gynecologic (GYN), and gastrointestinal (GI) cancer.

Clinical Characteristic	Breast	Sarcoma	H&N	GYN	GI
Sex	Female	Female	Male	Female	Male
T-stage	T2	T1	T3	T3b	T2
N-stage	N1	N0	N1	N1	N1c
M-stage	M0	M0	M0	M0	M0
Misc.*	ER-/PR-/HER2-	FNCLCC grade 3	HPV+	FIGO Stage IIIC1	Moderately differentiated

* Miscellaneous clinical characteristics specific to each cancer type are provided. ER-/PR-/HER2- = triple negative breast cancer. FNCLCC = Fédération Nationale des Centres de Lutte Contre le Cancer. HPV = human papillomavirus. FIGO = International Federation of Gynaecology and Obstetrics.

Imaging Protocols. Each patient received a radiotherapy planning CT scan which was exported in Digital Imaging and Communications in Medicine (DICOM) format. Image acquisition characteristics are shown in **Table 2**. The sarcoma, H&N, and GI cases received intravenous contrast, the GU case received oral contrast, and the breast case did not receive any contrast. Of note, the H&N case had metal streak artifacts secondary to metallic implants in the upper teeth, which obscured anatomy near the mandible. No other cases contained noticeable image artifacts.

Table 2. Computed tomography (CT) acquisition parameters for cases included in this data descriptor. Cases included breast, sarcoma, head and neck (H&N), gynecologic (GYN), and gastrointestinal (GI) cancer.

CT Acquisition Parameter	Breast	Sarcoma	H&N	GYN	GI

Manufacturer	GE	SIEMENS	SIEMENS	GE	GE
Model	Discovery CT590 RT	SOMATOM Confidence	Sensation Open	Discovery CT590 RT	Discovery CT590 RT
Slice Thickness (mm)	2.5	3	3	2.5	2.5
KVP (kV)	120	120	120	120	120
Exposure Time (ms)	891	1000	1000	856	856
X-Ray Tube Current (mA)	154	111	32	167	277
Rows	512	512	512	512	512
Columns	512	512	512	512	512
In-plane Resolution (mm)	1.26	1.26	0.98	0.98	0.98
Reconstruction diameter (mm)	650	650	500	500	500
Number of axial slices	140	229	143	195	196

Participant details. To register for the challenge, participants completed a baseline questionnaire that included their name, email address, affiliated institution, country, specialization, years in practice, number of disease sites treated, volume of patients treated per month for the designated tumor site, how they learned about this challenge, and reasons for participation. The research conducted herein was approved by the Human Research Protection Program (HRPP) at MSK (IRB#: X19-040 A(1); approval date: May 26, 2021). All subjects prospectively consented to participation in the present study, as well as to the collection, use, and disclosure of de-identified aggregate subject information and responses. Participants were categorized as recognized experts or non-experts. Recognized experts were identified by our C3RO team (EFG, CDF, DL) based on participation in the development of national guidelines or other extensive scholarly activities. Recognized experts were board-certified physicians with expertise in the specific disease site. Out of 1,026 registrants, 221 participated in generating segmentations, which were used for this dataset. Of note, participants could provide segmentations for multiple cases. Additional demographic characteristics of the participants can be found in Lin & Wahid et al. [19].

ProKnow segmentation platform. Participants were given access to the C3RO workspace on ProKnow (Elekta AB, Stockholm, Sweden). ProKnow is a cloud-based segmentation platform that stores and manages radiotherapy data. Anonymized CT DICOM images for each case were imported into the ProKnow system for participants to segment. Each case was attributed a short text prompt describing the patient presentation along with any additional information as needed. Participants were allowed to utilize common image manipulation (scrolling capabilities, zooming capabilities, window leveling, etc.) and segmentation (fill, erase, etc.) tools for generating their segmentations. No auto-segmentation capabilities were provided to the participants, i.e., all segmentations were manually generated. Notably, for the sarcoma case, an external mask of the patient's body and a mask of the left femur was provided to participants. Screenshots of the ProKnow web interface platform for the various cases are shown in **Fig. 2**.

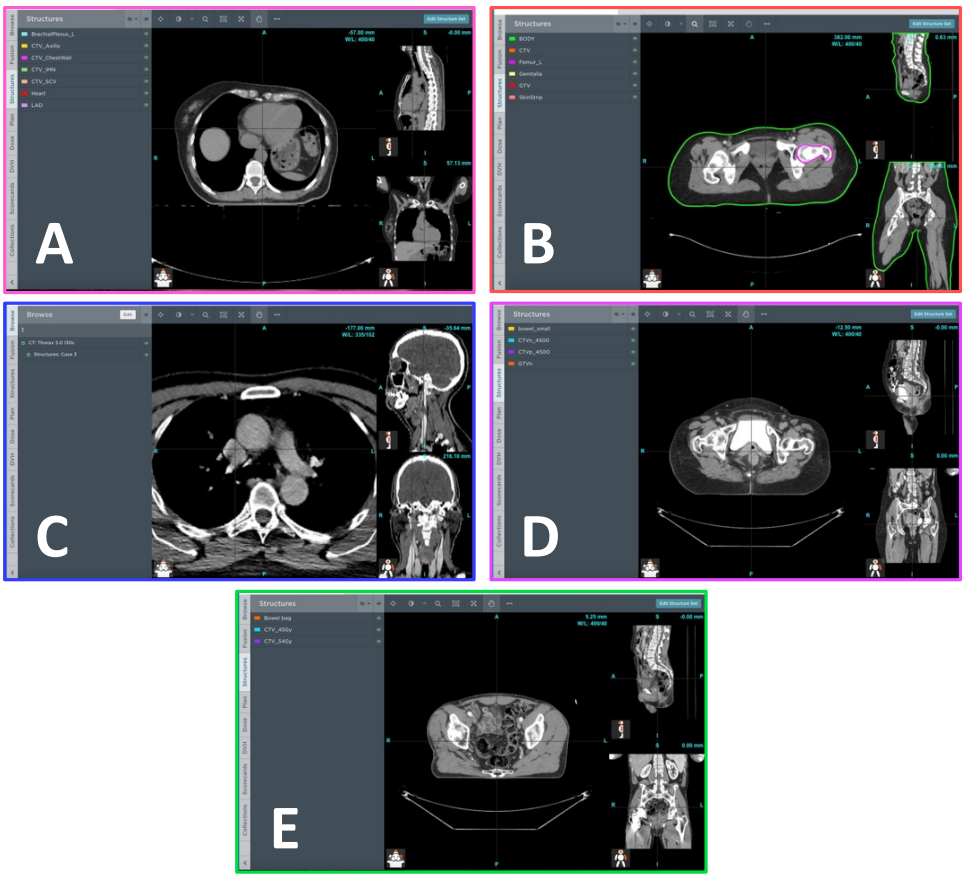

Figure 2. Examples of ProKnow segmentation platform used by participants for the breast **(A)**, sarcoma **(B)**, head and neck **(C)**, gynecologic **(D)**, and gastrointestinal **(E)** cases. Participants were given access to the standard image visualization and segmentation capabilities for generating their segmentations of target structures and organs at risk. Participants were also given access to a short prompt describing the patient presentation along with any additional information as needed. For the sarcoma case, an external mask of the patient's body (green) and a mask of the left femur (pink) was provided to participants. Subplots for breast, sarcoma,

head and neck, gynecologic, and gastrointestinal cases are outlined in pink, red, blue, purple, and green borders, respectively.

Segmentation details. For each case, participants were requested to segment a select number of ROIs corresponding to target structures or OARs. Notably, not all participants generated segmentations for all ROIs. ROIs for each participant were combined into one structure set in the ProKnow system. ROIs were initially named in a consistent, but non-standardized format, so during file conversion ROIs were renamed based on the The Report of American Association of Physicists in Medicine Task Group 263 (TG-263) suggested nomenclature [20]; TG-263 was chosen due its ubiquity in standardized radiotherapy nomenclature. A list of each ROI and the number of available segmentations stratified by participant expertise level is shown in **Table 3**.

Table 3. Summary of all region of interest (ROI) segmentations generated by participants for this data descriptor. ROIs included radiotherapy target volumes and organs at risk (OARs).

Case	Type of ROI	ROI	Definition(s)	Number of expert segmentations *	Number of non-expert segmentations *
Breast	Target volumes	BrachialPlex_L	Brachial plexus left	6	88
		CTV_Ax	Clinical target volume of axillary region	8	115
		CTV_Chestwall	Clinical target volume of chest wall	8	117
		CTV_IMN	Clinical target volume of internal mammary nodes	8	118
		CTV_Sclav_LN	Clinical target volume of supraclavicular lymph nodes	8	119
	OARs	Heart	Heart	7	121
		A_LAD_L	Left anterior descending artery	7	88
Sarcoma	Target volumes	GTV	Gross tumor volume	5	60
		CTV	Clinical tumor volume	5	48
	OARs	Genitals	Genitalia	4	51
		GTVp	Gross tumor volume primary -	14	59

Head and Neck	Target volumes		right tonsillar fossa		
		GTVn	Gross tumor volume of nodes - nodal spread to level II/III on ipsilateral side (with sternocleidomastoid muscle invaded) and no contralateral nodal involvement	13	60
		CTV1	Clinical target volume (high-risk)	9	45
		CTV2	Clinical target volume (low to intermediate risk)	9	49
	OARs	Brainstem	Brainstem	13	58
		Glnd_Submand_L	Submandibular gland left	13	57
		Glnd_Submand_R	Submandibular gland right	12	52
		Larynx	Larynx	12	57
		Musc_Constrict	All pharyngeal constrictor muscles (superior, middle, and inferior)	11	43
		Partoid_L	Parotid left	13	59
		Parotid_R	Parotid right	13	58
Gynecologic	Target volumes	GTVn	Gross tumor volume of the involved right common iliac lymph node	5	42
		CTVn_4500	Clinical target volume for the elective nodal volumes at risk that will receive 45 Gy	5	40
		CTVp_4500	Clinical target volume primary will receive 45 Gy. This is the combination of	5	41

			"ctv1" and "ctv2" used in many RTOG protocols		
	OARs	Bowel_Small	Small bowel	4	35
Gastrointestinal	Target volumes	CTV_4500	Clinical target volume that will receive 45 Gy	4	25
		CTV_5400	Clinical target volume that will receive 54 Gy	4	23
	OARs	Bag_Bowel	Small and large bowel	4	23

* Not all participants generated segmentations for all ROIs.

Image processing and file conversion. For each case, anonymized CT images and structure sets for each annotator were manually exported from ProKnow in DICOM and DICOM radiotherapy structure (RTS) format, respectively. The Neuroimaging Informatics Technology Initiative (NIfTI) format is increasingly seen as an important standard for reproducible imaging research due to its compact file size and ease of implementation in computational models [21]. Therefore, in order to increase the interoperability of these data, we converted all our DICOM imaging and segmentation data to NIfTI format. For all file conversion processes, Python v. 3.8.8 [22] was used. An overview of the image processing workflow is shown in **Fig. 3A**. In brief, using an in-house Python script, DICOM images and structure sets were loaded into numpy array format using the DICOMRTTool v. 0.4.2 library [23], and then converted to NIfTI format using SimpleITK v. 2.1.1 [24]. For each annotator, each individual structure contained in the structure set was separately converted into a binary mask (0 = background, 1 = ROI), and was then individually converted into separate NIfTI files. Examples of random subsets of five expert segmentations for each ROI from each case are shown in **Fig. 4**.

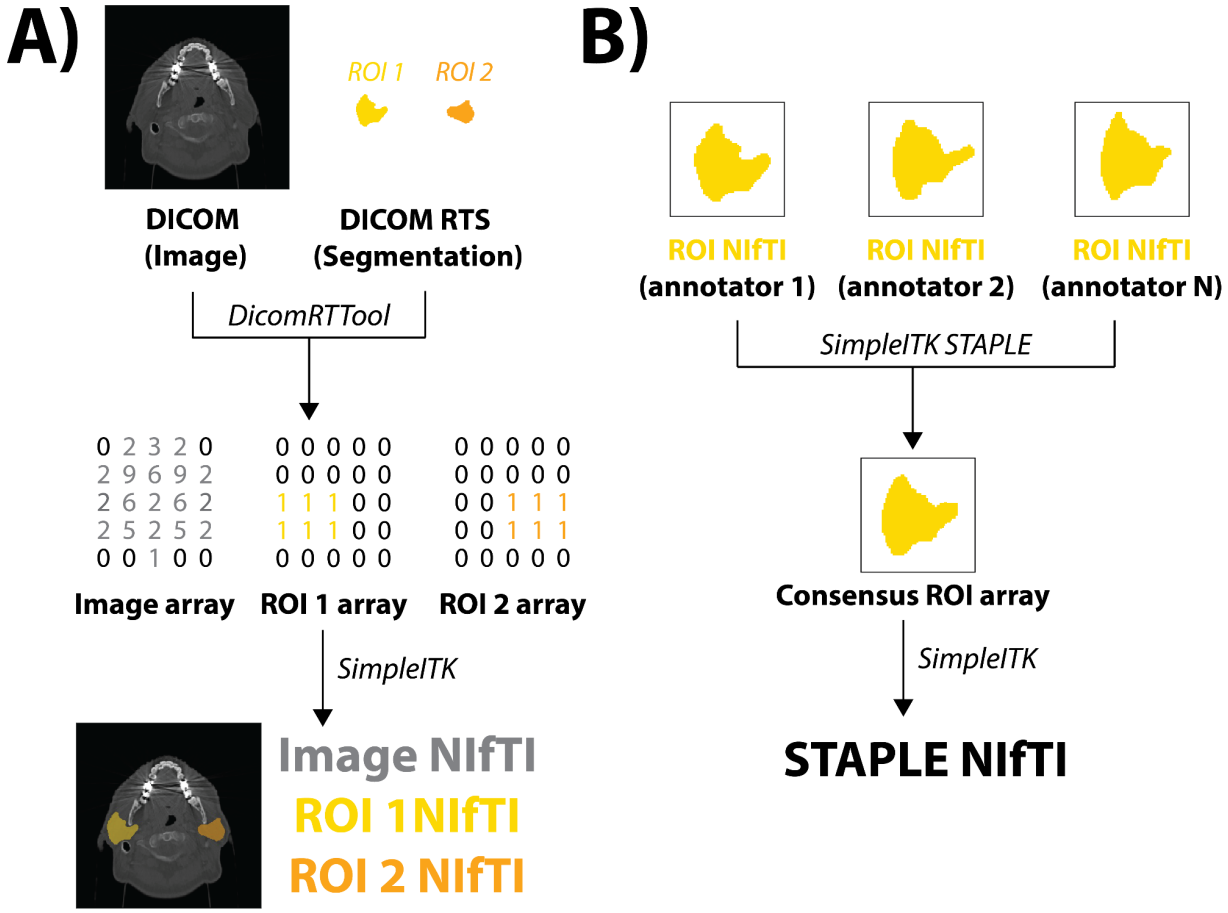

Figure 3. Image processing Python workflows implemented for this data descriptor for an example case (head and neck case, parotid glands). **(A)** Original Digital Imaging and Communications in Medicine (DICOM) formatted images and DICOM radiotherapy structure (RTS) formatted region of interest (ROI) segmentations are transformed to Neuroimaging Informatics Technology Initiative (NIfTI) format. **(B)**. ROIs from multiple annotators are combined into a single consensus segmentation using the Simultaneous truth and performance level estimation (STAPLE) method.

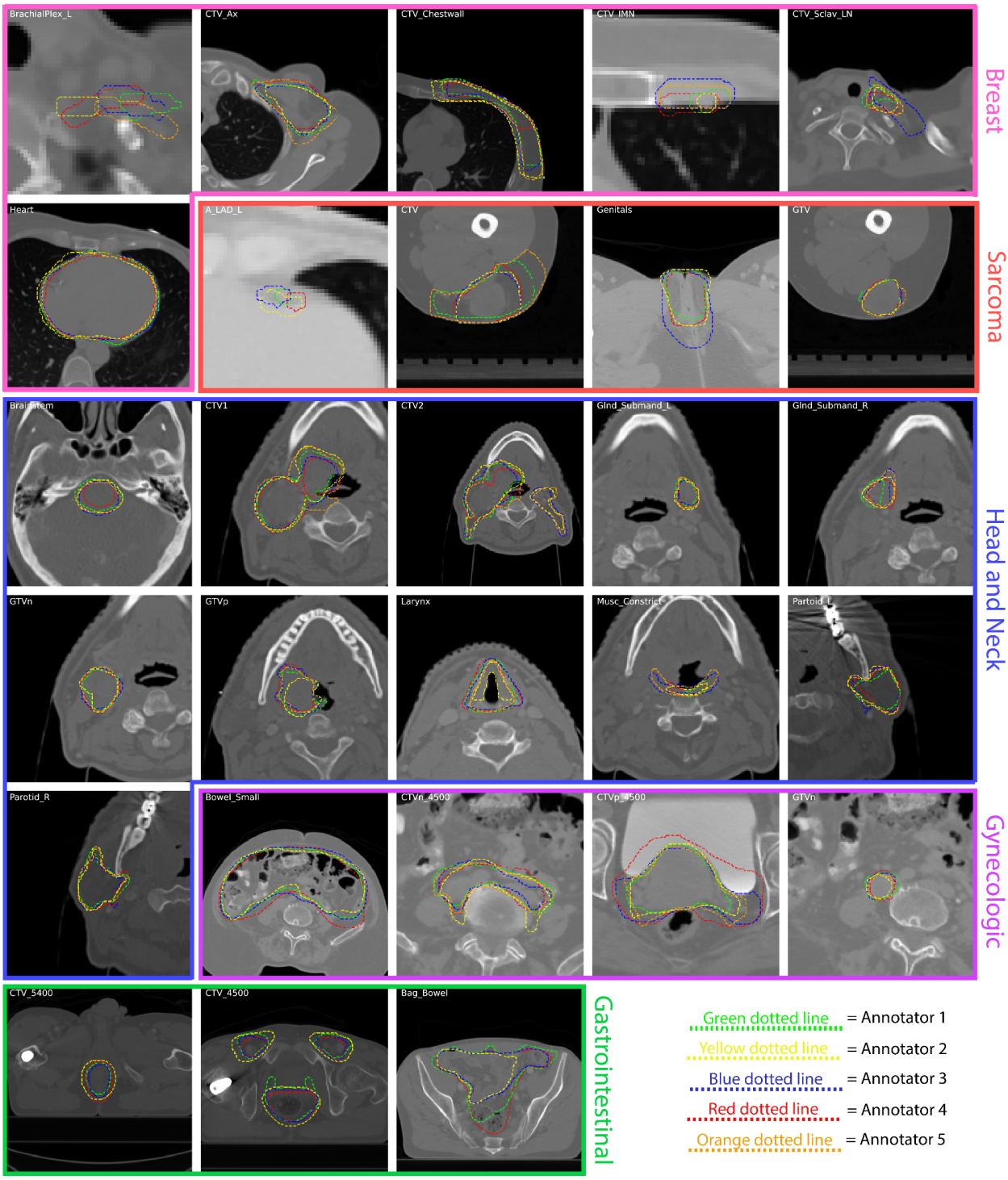

Figure 4. Examples of a random subset of five expert segmentations for each region of interest (ROI) provided in this data descriptor. Segmentations are displayed as orange, yellow, blue, red, and green dotted lines corresponding to annotators 1, 2, 3, 4, and 5, respectively, and overlaid on zoomed-in images for each case. Subplots for breast, sarcoma, head and neck, gynecologic, and gastrointestinal cases are outlined in pink, red, blue, purple, and green borders, respectively. Notably, the gastrointestinal case only had four annotators, so only four lines are displayed.

Consensus segmentation generation. In addition to ground-truth expert and non-expert segmentations for all ROIs, we also generated consensus segmentations using the simultaneous truth and performance level estimation (STAPLE) method, a commonly used probabilistic approach for combining multiple segmentations [25–28]. Briefly, the STAPLE method uses an iterative expectation-maximization algorithm to compute a probabilistic estimate of the "true" segmentation by deducing an optimal combination of the input segmentations and incorporating a prior model for the spatial distribution of segmentations as well as implementing spatial homogeneity constraints [29]. For our specific implementation of the STAPLE method, we utilized the SimpleITK STAPLE function with a default threshold value of 0.95. For each ROI, all available binary segmentation masks acted as inputs to the STAPLE function for each expertise level, subsequently generating binary STAPLE segmentation masks for each expertise level (i.e., $STAPLE_{expert}$ and $STAPLE_{non\text{-}expert}$). An overview of the consensus segmentation workflow is shown in **Fig. 3B**. Examples of $STAPLE_{expert}$ and $STAPLE_{non\text{-}expert}$ segmentations for each ROI are shown in **Fig. 5**.

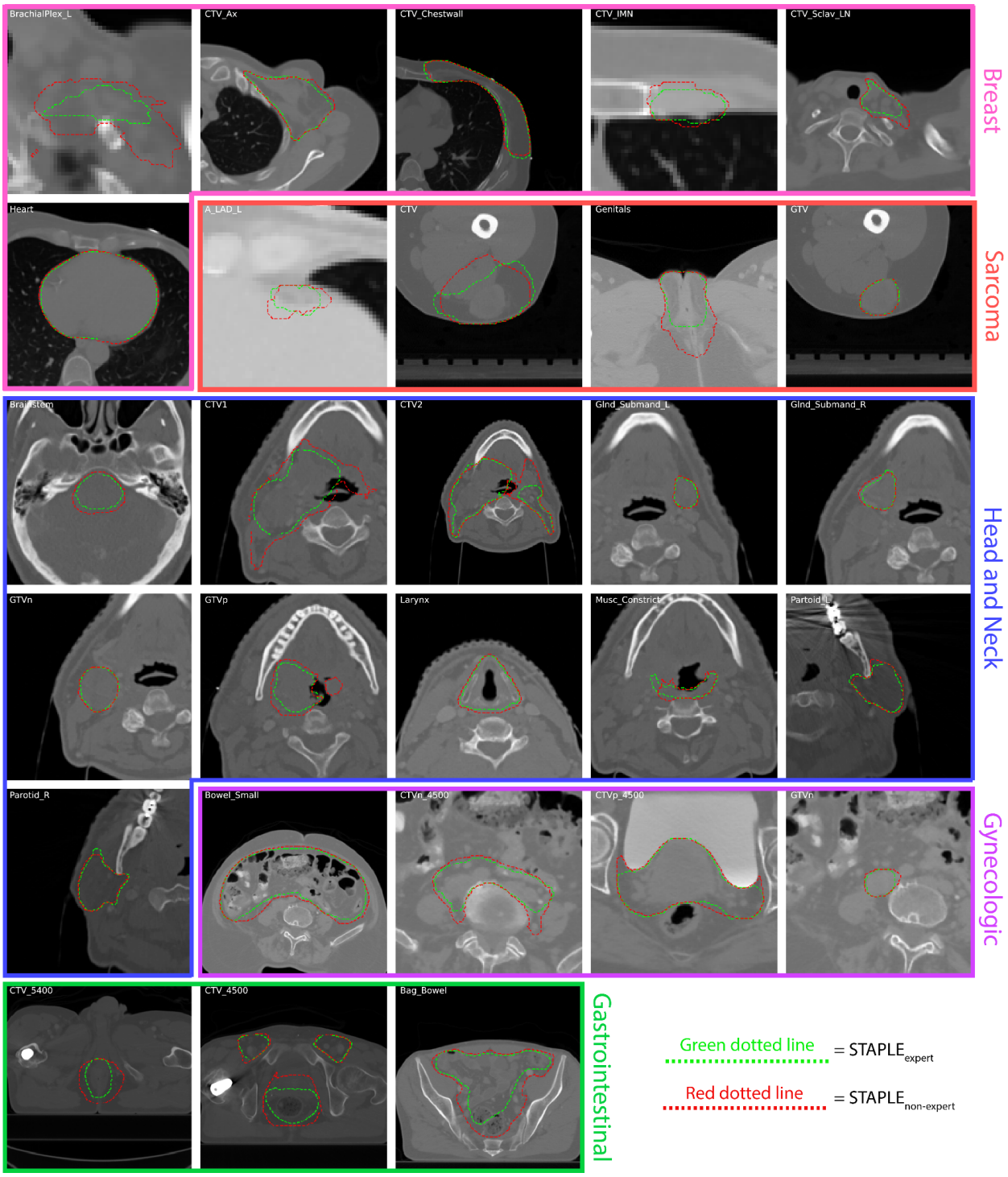

Figure 5. Examples of consensus segmentations using the simultaneous truth and performance level estimation (STAPLE) method for each region of interest (ROI) provided in this data descriptor. STAPLE segmentation generated by using all available expert segmentations ($STAPLE_{expert}$) and STAPLE segmentation generated by using all available non-expert segmentations ($STAPLE_{non\text{-}expert}$) are displayed as green and red dotted outlines, respectively, and overlaid on zoomed in images for each case. Subplots for breast, sarcoma, head and neck,

gynecologic, and gastrointestinal cases are outlined in pink, red, blue, purple, and green borders, respectively.

Data Records

CT images and multi-annotator segmentation data. This data collection primarily consists of 1985 3D volumetric compressed NIfTI files (.nii.gz file extension) corresponding to CT images and segmentations of ROIs from various disease sites (breast, sarcoma, H&N, GYN, GI). ROI segmentation NIfTI files are provided in binary mask format (0 = background,1 = ROI); file names for each ROI are provided in TG-263 notation. All CT images and ROI segmentations were derived from original DICOM and DICOM RTS files (.dcm file extension) respectively, which for completeness are also provided in this data collection. In addition, Python code to recreate the final NIfTI files from DICOM files is also provided in the corresponding GitHub repository (see *Code Availability* section).

Consensus segmentation data. Consensus segmentations for experts and non-experts generated using the STAPLE method for each ROI have also been provided in compressed NIfTI file format (.nii.gz file extension). Consensus segmentation NIfTI files are provided in binary mask format (0 = background, 1 = ROI consensus). Python code to recreate the STAPLE NIfTI files from input annotator NIfTI files is also provided in the corresponding GitHub repository (see *Code Availability* section).

Annotator demographics data. We also provide a single excel file (.xlsx file extension) containing each annotator's gender, race/ethnicity, geographic setting, profession, years of experience, practice type, and categorized expertise level (expert, non-expert). Geographic setting was re-coded as “United States” or “International” to further de-identify the data. Each separate sheet corresponds to a separate disease site (sheet 1 = breast, sheet 2 = sarcoma, sheet 3 = H&N, sheet 4 = GU, sheet 5 = GI).

Folder structure and identifiers. Each disease site is represented by a top-level folder, containing a subfolder for images and segmentations. The annotator demographic excel file is located in the same top-level location as the disease site folders. Image folders contain separate subfolders for NIfTI format and DICOM format images. Segmentation folders contain separate subfolders for expert and non-expert segmentations. Each expertise folder contains separate subfolders for each annotator (which contains separate subfolders for DICOM and NIfTI formatted files) and the consensus segmentation (only available in NIfTI format). An overview of the organized data records for an example case is shown in **Fig. 6**. Segmentation files (DICOM and NIfTI) are organized by anonymized participant ID numbers and can be cross referenced against the excel data table using this identifier. The raw data, records, and supplemental descriptions of the meta-data files are cited under Figshare doi: 10.6084/m9.figshare.21074182 (private until manuscript acceptance).

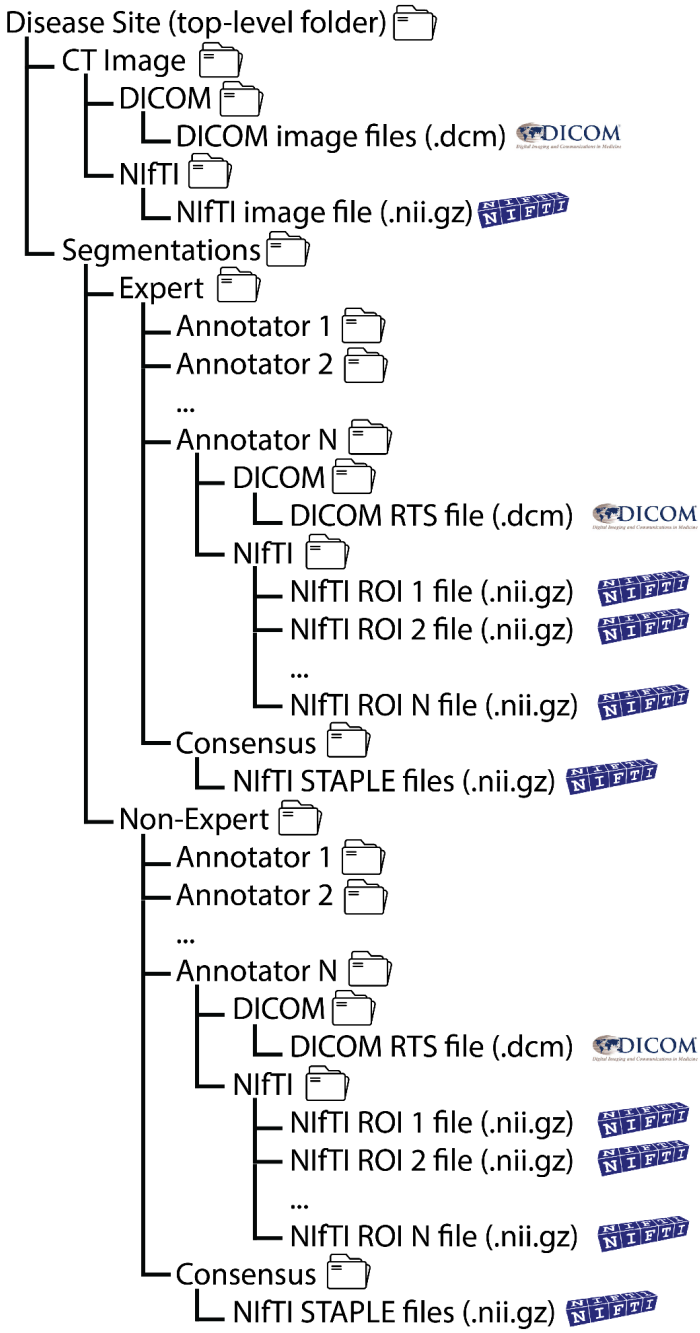

Figure 6. Overview of folder and file structure for dataset provided in this data descriptor. Each disease site folder contains separate subfolders for the computed tomography (CT) image and segmentations. Image subfolders contain separate subfolders for different data formats (Digital Imaging and Communications in Medicine [DICOM] and Neuroimaging Informatics Technology Initiative [NIfTI]). Segmentation subfolders contain separate subfolders which stratify expert and non-experts, which are further divided into subfolders for each annotator's segmented ROIs in DICOM and NIfTI formats. Consensus segmentations for each ROI generated by the simultaneous truth and performance level estimation (STAPLE) method are also provided in expert and non-expert folders.

Technical Validation

Elekta ProKnow platform. ProKnow is a commercially available radiotherapy clinical workflow tool that allows for centralization of data in a secure web-based repository. The ProKnow system has been adopted by several large scale medical institutions and is used routinely in clinical and research environments.

Segmentation data. The segmentations provided in this data descriptor have been utilized in a study by Lin & Wahid et al. [19]. This study demonstrated several results that were consistent with existing literature, including: 1). target ROIs tended to exhibit greater variability than OAR ROIs [28], 2). H&N ROIs exhibited higher interobserver variability compared to other disease sites [30,31], and 3). non-expert consensus segmentations could approximate gold-standard expert segmentations [32].

Usage Notes

The image and segmentation data from this data collection are provided in original DICOM format (where applicable) and compressed NIfTI format with the accompanying excel file containing demographic information indexed by annotator identifiers. We invite all interested researchers to download this dataset for use in segmentation, radiotherapy, and crowdsourcing related research. Moreover, we encourage this dataset's use for clinical decision support tool development. Importantly, this dataset could also be used as a standardized reference for future interobserver variability studies seeking to investigate further participant expertise criteria, e.g., true novice annotators (no previous segmentation or anatomy knowledge) could attempt to segment ROI structures on CT images, which could then be compared to our expert and non-expert annotators. Finally, in line with the goals of the eContour collaborative [33], these data could be used to help develop educational tools for radiation oncology clinical training.

Original DICOM format images and structure sets may be viewed and analyzed in radiation treatment planning software or select digital image viewing applications, depending on the end-user's requirements. Current commercial treatment planning software for these purposes include Velocity AI (Varian Medical Systems; Palo Alto, United States), Raystation [34], and MIM Maestro [35]. Current open-source software for these purposes includes ImageJ [36], dicompyler [37], ITK-Snap [38], and 3D Slicer [39] with the SlicerRT extension [40].

Processed NIfTI format images and segmentations may be viewed and analyzed in any NIfTI viewing application, depending on the end-user's requirements. Current open-source software for these purposes includes ImageJ [36], ITK-Snap [38], and 3D Slicer [39].

Code Availability

Segmentations were performed using the commercially-available ProKnow (Elekta AB, Stockholm, Sweden) software. The code for NIfTI file conversion of DICOM CT images and corresponding DICOM RTS segmentations, along with code for consensus segmentation

generation, was developed using in-house Python scripts and is made publicly available through GitHub: https://github.com/kwahid/C3RO_analysis (private until manuscript acceptance).

References

1. Sharp, G., Fritscher, K. D., Pekar, V., Peroni, M., Shusharina, N., Veeraraghavan, H. & Yang, J. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med. Phys.* **41,** 050902 (2014).
2. Segedin, B. & Petric, P. Uncertainties in target volume delineation in radiotherapy – are they relevant and what can we do about them? *Radiology and Oncology* **50,** 254–262 Preprint at https://doi.org/10.1515/raon-2016-0023 (2016)
3. Njeh, C. F. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *J. Med. Phys.* **33,** 136–140 (2008).
4. Berry, S. L., Boczkowski, A., Ma, R., Mechalakos, J. & Hunt, M. Interobserver variability in radiation therapy plan output: Results of a single-institution study. *Practical Radiation Oncology* **6,** 442–449 Preprint at https://doi.org/10.1016/j.prro.2016.04.005 (2016)
5. Sherer, M. V., Lin, D., Elguindi, S., Duke, S., Tan, L.-T., Cacicedo, J., Dahele, M. & Gillespie, E. F. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiother. Oncol.* **160,** 185–191 (2021).
6. Cardenas, C. E., Yang, J., Anderson, B. M., Court, L. E. & Brock, K. B. Advances in Auto-Segmentation. *Semin. Radiat. Oncol.* **29,** 185–197 (2019).
7. Harrison, K., Pullen, H., Welsh, C., Oktay, O., Alvarez-Valle, J. & Jena, R. Machine Learning for Auto-Segmentation in Radiotherapy Planning. *Clin. Oncol.* **34,** 74–88 (2022).
8. Yu, S., Chen, M., Zhang, E., Wu, J., Yu, H., Yang, Z., Ma, L., Gu, X. & Lu, W. Robustness study of noisy annotation in deep learning based medical image segmentation. *Phys. Med. Biol.* **65,** 175007 (2020).
9. Budd, S., Day, T., Simpson, J., Lloyd, K., Matthew, J., Skelton, E., Razavi, R. & Kainz, B. Can Non-specialists Provide High Quality Gold Standard Labels in Challenging Modalities?

in *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health* 251–262 (Springer International Publishing, 2021).

10. Heim, E., Roß, T., Seitel, A., März, K., Stieltjes, B., Eisenmann, M., Lebert, J., Metzger, J., Sommer, G., Sauter, A. W., Schwartz, F. R., Termer, A., Wagner, F., Kenngott, H. G. & Maier-Hein, L. Large-scale medical image annotation with crowd-powered algorithms. *J Med Imaging (Bellingham)* **5,** 034002 (2018).
11. Wesemeyer, T., Jauer, M.-L. & Deserno, T. M. Annotation quality vs. quantity for deep-learned medical image segmentation. in *Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications* **11601,** 63–76 (SPIE, 2021).
12. Afshar, P., Heidarian, S., Enshaei, N., Naderkhani, F., Rafiee, M. J., Oikonomou, A., Fard, F. B., Samimi, K., Plataniotis, K. N. & Mohammadi, A. COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. *Sci Data* **8,** 121 (2021).
13. Wallner, J., Mischak, I. & Jan Egger. Computed tomography data collection of the complete human mandible and valid clinical ground truth models. *Sci Data* **6,** 190003 (2019).
14. Lyu, X., Cheng, L. & Zhang, S. The RETA Benchmark for Retinal Vascular Tree Analysis. *Sci Data* **9,** 397 (2022).
15. Kulaga-Yoskovitz, J., Bernhardt, B. C., Hong, S.-J., Mansi, T., Liang, K. E., van der Kouwe, A. J. W., Smallwood, J., Bernasconi, A. & Bernasconi, N. Multi-contrast submillimetric 3 Tesla hippocampal subfield segmentation protocol and dataset. *Sci Data* **2,** 150059 (2015).
16. Payette, K., de Dumast, P., Kebiri, H., Ezhov, I., Paetzold, J. C., Shit, S., Iqbal, A., Khan, R., Kottke, R., Grehten, P., Ji, H., Lanczi, L., Nagy, M., Beresova, M., Nguyen, T. D., Natalucci, G., Karayannis, T., Menze, B., Bach Cuadra, M. & Jakab, A. An automatic multi-tissue human fetal brain segmentation benchmark using the Fetal Tissue Annotation Dataset. *Sci Data* **8,** 167 (2021).
17. Jin, K., Huang, X., Zhou, J., Li, Y., Yan, Y., Sun, Y., Zhang, Q., Wang, Y. & Ye, J. FIVES: A

Fundus Image Dataset for Artificial Intelligence based Vessel Segmentation. *Sci Data* **9,** 475 (2022).

18. Lin, D., Fuller, D., Nelms, B. E., Mohamed, A. S., Duke, S., Sherer, M. V., Cislo, M., Murphy, J. & Gillespie, E. F. Contouring Collaborative for Consensus in Radiation Oncology (C3RO): An International Crowdsourcing Challenge to Improve Radiotherapy Contour Delineation. *Int. J. Radiat. Oncol. Biol. Phys.* **111,** e10–e11 (2021).
19. Lin, D., Wahid, K. A., Nelms, B. E., He, R., Naser, M. A., Duke, S., Sherer, M. V., Christodouleas, J. P., Mohamed, A. S. R., Cislo, M., Murphy, J. D., Fuller, C. D. & Gillespie, E. F. 'E pluribus unum': Prospective acceptability benchmarking from the contouring collaborative for consensus in Radiation Oncology (C3RO) crowdsourced initiative for multi-observer segmentation. *medRxiv* (2022). doi:10.1101/2022.09.23.22280295
20. Mayo, C. S., Moran, J. M., Bosch, W., Xiao, Y., McNutt, T., Popple, R., Michalski, J., Feng, M., Marks, L. B., Fuller, C. D., Yorke, E., Palta, J., Gabriel, P. E., Molineu, A., Matuszak, M. M., Covington, E., Masi, K., Richardson, S. L., Ritter, T., Morgas, T., Flampouri, S., Santanam, L., Moore, J. A., Purdie, T. G., Miller, R. C., Hurkmans, C., Adams, J., Jackie Wu, Q.-R., Fox, C. J., Siochi, R. A., Brown, N. L., Verbakel, W., Archambault, Y., Chmura, S. J., Dekker, A. L., Eagle, D. G., Fitzgerald, T. J., Hong, T., Kapoor, R., Lansing, B., Jolly, S., Napolitano, M. E., Percy, J., Rose, M. S., Siddiqui, S., Schadt, C., Simon, W. E., Straube, W. L., St James, S. T., Ulin, K., Yom, S. S. & Yock, T. I. American Association of Physicists in Medicine Task Group 263: Standardizing Nomenclatures in Radiation Oncology. *Int. J. Radiat. Oncol. Biol. Phys.* **100,** 1057–1066 (2018).
21. Li, X., Morgan, P. S., Ashburner, J., Smith, J. & Rorden, C. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *J. Neurosci. Methods* **264,** 47–56 (2016).
22. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, 2009).
23. Anderson, B. M., Wahid, K. A. & Brock, K. K. Simple Python Module for Conversions Between DICOM Images and Radiation Therapy Structures, Masks, and Prediction Arrays.

Pract. Radiat. Oncol. **11,** 226–229 (2021).

24. Yaniv, Z., Lowekamp, B. C., Johnson, H. J. & Beare, R. SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. *J. Digit. Imaging* **31,** 290–303 (2018).
25. Taku, N., Wahid, K. A., van Dijk, L. V., Sahlsten, J., Jaskari, J., Kaski, K., Fuller, C. D. & Naser, M. A. Auto-detection and segmentation of involved lymph nodes in HPV-associated oropharyngeal cancer using a convolutional deep learning neural network. *Clinical and Translational Radiation Oncology* **36,** 47–55 (2022).
26. Naser, M. A., Wahid, K. A., van Dijk, L. V., He, R., Abdelaal, M. A., Dede, C., Mohamed, A. S. R. & Fuller, C. D. Head and Neck Cancer Primary Tumor Auto Segmentation Using Model Ensembling of Deep Learning in PET/CT Images. in *Head and Neck Tumor Segmentation and Outcome Prediction* 121–133 (Springer International Publishing, 2022).
27. McDonald, B. A., Cardenas, C., O'Connell, N., Ahmed, S., Naser, M. A., Wahid, K. A., Xu, J., Thill, D., Zuhour, R., Mesko, S., Augustyn, A., Buszek, S. M., Grant, S., Chapman, B. V., Bagley, A., He, R., Mohamed, A., Christodouleas, J. P., Brock, K. K. & Fuller, C. D. Investigation of autosegmentation techniques on T2-weighted MRI for off-line dose reconstruction in MR-linac Adapt to Position workflow for head and neck cancers. *bioRxiv* (2021). doi:10.1101/2021.09.30.21264327
28. Cardenas, C. E., Blinde, S. E., Mohamed, A. S. R., Ng, S. P., Raaijmakers, C., Philippens, M., Kotte, A., Al-Mamgani, A. A., Karam, I., Thomson, D. J., Robbins, J., Newbold, K., Fuller, C. D. & Terhaard, C. Comprehensive Quantitative Evaluation of Variability in Magnetic Resonance-Guided Delineation of Oropharyngeal Gross Tumor Volumes and High-Risk Clinical Target Volumes: An R-IDEAL Stage 0 Prospective Study. *Int. J. Radiat. Oncol. Biol. Phys.* **113,** 426–436 (2022).
29. Warfield, S. K., Zou, K. H. & Wells, W. M. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans.*

Med. Imaging **23,** 903–921 (2004).

30. van der Veen, J., Gulyban, A., Willems, S., Maes, F. & Nuyts, S. Interobserver variability in organ at risk delineation in head and neck cancer. *Radiat. Oncol.* **16,** 120 (2021).
31. Brouwer, C. L., Steenbakkers, R. J. H. M., van den Heuvel, E., Duppen, J. C., Navran, A., Bijl, H. P., Chouvalova, O., Burlage, F. R., Meertens, H., Langendijk, J. A. & van ’t Veld, A. A. 3D Variation in delineation of head and neck organs at risk. *Radiat. Oncol.* **7,** 32 (2012).
32. O’Neil, A. Q., Murchison, J. T., van Beek, E. J. R. & Goatman, K. A. Crowdsourcing Labels for Pathological Patterns in CT Lung Scans: Can Non-experts Contribute Expert-Quality Ground Truth? in *Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* 96–105 (Springer International Publishing, 2017).
33. Sherer, M. V., Lin, D., Puri, K., Panjwani, N., Zhang, Z., Murphy, J. D. & Gillespie, E. F. Development and Usage of eContour, a Novel, Three-Dimensional, Image-Based Web Site to Facilitate Access to Contouring Guidelines at the Point of Care. *JCO Clin Cancer Inform* **3,** 1–9 (2019).
34. Bodensteiner, D. RayStation: External beam treatment planning system. *Med. Dosim.* **43,** 168–176 (2018).
35. Brinkmann & Piper. Image registration, deformation, and enhanced contouring for radiotherapy with MIM Maestro. *Mayo Clinic Rochester, MN and MIM Software* at <https://www.aapm.org/meetings/amos2/pdf/60-14841-24267-158.pdf>
36. Abramoff, Magalhães & Ram. Image processing with ImageJ. *Biophotonics int.* at <https://dspace.library.uu.nl/handle/1874/204900>
37. Panchal, A. & Keyes, R. SU-GG-T-260: dicompyler: an open source radiation therapy research platform with a plugin architecture. *Med. Phys.* **37,** 3245–3245 (2010).
38. Yushkevich, P. A., Yang Gao & Gerig, G. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*

2016, 3342–3345 (2016).

39. Pieper, S., Halle, M. & Kikinis, R. 3D Slicer. in *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)* 632–635 Vol. 1 (2004).
40. Pinter, C., Lasso, A., Wang, A., Jaffray, D. & Fichtinger, G. SlicerRT: radiation therapy research toolkit for 3D Slicer. *Med. Phys.* **39,** 6332–6338 (2012).