scientific data

Check for updates

OPEN

DATA DESCRIPTOR

# Muscle and adipose tissue segmentations at the third cervical vertebral level in patients with head and neck cancer

Kareem A. Wahid[1], Brennan Olson[2,3], Rishab Jain[2], Aaron J. Grossberg[2], Dina El-Habashy[1,4], Cem Dede[1], Vivian Salama[1], Moamen Abobakr[1], Abdallah S. R. Mohamed[1], Renjie He[1], Joel Jaskari[5], Jaakko Sahlsten[5], Kimmo Kaski[5], Clifton D. Fuller[1] ✉ & Mohamed A. Naser[1] ✉

**The accurate determination of sarcopenia is critical for disease management in patients with head and neck cancer (HNC). Quantitative determination of sarcopenia is currently dependent on manually-generated segmentations of skeletal muscle derived from computed tomography (CT) cross-sectional imaging. This has prompted the increasing utilization of machine learning models for automated sarcopenia determination. However, extant datasets currently do not provide the necessary manually-generated skeletal muscle segmentations at the C3 vertebral level needed for building these models. In this data descriptor, a set of 394 HNC patients were selected from The Cancer Imaging Archive, and their skeletal muscle and adipose tissue was manually segmented at the C3 vertebral level using sliceOmatic. Subsequently, using publicly disseminated Python scripts, we generated corresponding segmentations files in Neuroimaging Informatics Technology Initiative format. In addition to segmentation data, additional clinical demographic data germane to body composition analysis have been retrospectively collected for these patients. These data are a valuable resource for studying sarcopenia and body composition analysis in patients with HNC.**

## Background & Summary

Head and neck cancer (HNC) affects more than 900,000 individuals worldwide annually[1]. Sarcopenia, a body composition status describing skeletal muscle depletion, is a well-validated negative prognostic factor in patients with HNC and has become increasingly studied in recent years[2–4]. Sarcopenia is quantitatively determined primarily using the cross-sectional estimate of skeletal muscle at a specific vertebral level. Current methods to generate cross-sectional skeletal muscle segmentations for use in sarcopenia determination are reliant on expert human-generated segmentations, which can be time-consuming to procure and subject to user variability[5]. Therefore, the dissemination of high-quality skeletal muscle segmentations is of paramount importance to develop tools for sarcopenia-related clinical decision making.

Publicly disseminated HNC datasets have increased sharply in recent years. For example, several HNC imaging datasets, predominantly composed of computed tomography (CT) images, have been hosted on The Cancer Imaging Archive (TCIA)[6]. Public datasets, such as these, have been crucial towards advanced algorithmic development for clinical decision support tools[7]. However, only a handful of existing HNC datasets have provided information germane to determining sarcopenia status in patients, namely that by Grossberg *et al.*[8] providing body composition analysis data based on abdominal imaging. Moreover, to date, there are no existing open-source repositories for body composition analysis data based on head and neck region imaging. Increasing

[1]Department of Radiation Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA. [2]Department of Radiation Medicine, Oregon Health & Science University, Portland, Oregon, USA. [3]Medical Scientist Training Program, Oregon Health & Science University, Portland, Oregon, USA. [4]Department of Clinical Oncology, Menoufia University, Shibin Al Kawm, Egypt. [5]Department of Computer Science, Aalto University School of Science, Espoo, Finland. ✉e-mail: cdfuller@mdanderson.org; manaser@mdanderson.org

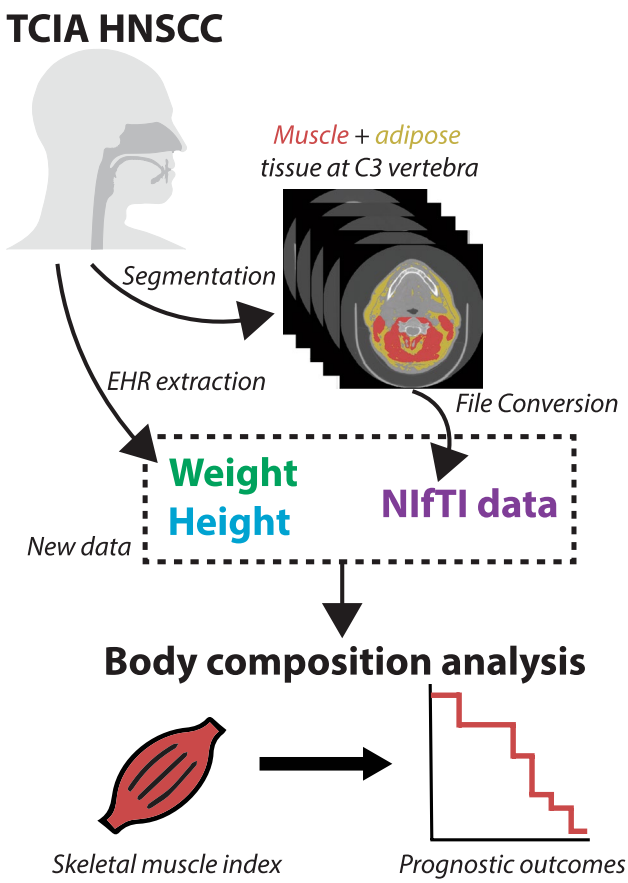


**Fig. 1** Data descriptor overview. The Cancer Imaging Archive (TCIA) head and neck squamous cell carcinoma (HNSCC) computed tomography dataset is used to generate muscle and adipose tissue segmentations at the third cervical (C3) vertebral level in Neuroimaging Informatics Technology Initiative (NIfTI) format. Additional demographic data (weight, height) is collected from electronic health records (EHR). The final newly distributed dataset can be used for body composition analysis, such as sarcopenia-related clinical decision-making.

evidence has shown the potential utility of sarcopenia determination using skeletal muscle in the head and neck region[2,9]. This is driven by the fact that many patients with HNC may not have abdominal imaging acquired as part of the standard workup, but will almost certainly have head and neck region imaging acquired, particularly due to its necessity for radiotherapy treatment planning[10] and staging purposes[11]. These head and neck imaging data could be used to train models for automated sarcopenia-related clinical decision making, as shown in previous studies[12]. Therefore, the dissemination of sarcopenia-related data derived from head and neck imaging is an unmet need that may foster more rapid adoption of automated HNC clinical decision support tools.

Here we present the curation and annotation of a large-scale TCIA dataset of 394 patients with HNC for use in sarcopenia-related clinical decision making and body composition analysis. The primary contribution of this dataset is high-quality skeletal muscle and adipose tissue segmentation at the cervical vertebral level in an easily accessible and standardized imaging format, in addition to additional clinical demographic variables. These data can be leveraged to build models for body composition analysis and sarcopenia-related decision-making germane to HNC. Moreover, these data could form the basis for future data modeling challenges for sarcopenia-related decision-making in patients with HNC. An overview of the data descriptor is shown in Fig. 1.

## Methods

**Study population and image details.** To develop this dataset, imaging data from the TCIA head and neck squamous cell carcinoma (HNSCC) collection, a large repository of imaging data originally collected from The University of Texas MD Anderson Cancer Center, were utilized. Specifically, 396 patients with contrast-enhanced CT scans were selected from the 495 available patients in the "Radiomics outcome prediction in Oropharyngeal cancer" dataset[13,14]. These patients were selected due to their inclusion of the third cervical vertebral level on imaging. To summarize the underlying data, these were patients with histopathologically-proven diagnosis of squamous cell carcinoma of the oropharynx that were treated with curative-intent intensity-modulated radiotherapy. Imaging data was composed of high-quality CT scans of patients who were injected with intravenous contrast material. Images were acquired before the start of radiotherapy. Imaging data were provided in the Digital

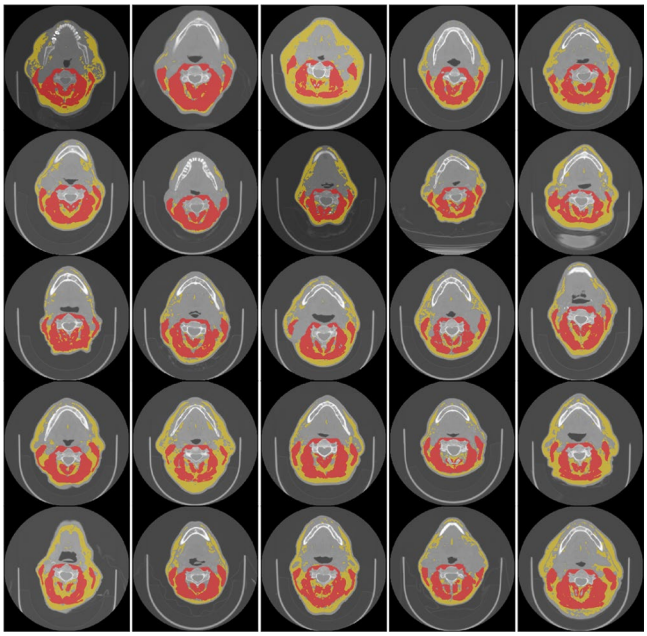

**Fig. 2** Segmentation examples for a subset of 25 cases. Each image corresponds to one patient. Images are single-slice computed tomography axial views with segmentations superimposed. The red regions correspond to skeletal muscle tissue and the yellow regions correspond to adipose tissue.

Imaging and Communications in Medicine (DICOM) standardized format. Additional details on the original imaging dataset are provided in the corresponding data descriptor[14] and TCIA website[13]. All DICOM images were previously de-identified, as described in previous data descriptors[8,14].

**Skeletal muscle segmentation.** For each CT image, the middle of the third cervical vertebra (C3) was located on a single axial slice and the skeletal muscle and adipose tissues were manually segmented. As described in previous publications[15], muscle and adipose tissue were defined in the ranges of −29 to 150 and −190 to −30 Hounsfield units, respectively to initially guide manual segmentation; manual corrections to the initial automatically generated segmentation were necessary due to the presence of non-desired tissues (i.e., vasculature, soft tissue) in the Hounsfield unit ranges implemented. Based on these criteria, the paraspinal and sternocleidomastoid muscles were included as part of the skeletal muscle segmentation, while subcutaneous, intermuscular, and visceral adipose compartments were included as part of the adipose segmentation. Skeletal muscle and adipose tissue were segmented by trained research assistants (B.O. and R.J.) and reviewed by a radiation oncologist with 4 years of post-residency experience (A.J.G.) using a commercial image-processing platform (sliceOmatic v. 5.0, Tomovision, Magog, Canada). Examples of skeletal muscle and adipose tissue segmentations with corresponding images are shown in Fig. 2. Segmentations were exported from sliceOmatic in .tag format, with the corresponding 2D axial slice in DICOM format.

**NIfTI conversion.** The Neuroimaging Informatics Technology Initiative (NIfTI) file format is increasingly seen as the standard for reproducible medical imaging research[16]. Therefore, we converted all our segmentation (.tag) and imaging (.dcm) data to NIfTI format, in order to increase the interoperability and widespread utilization of these data.

For all file conversion processes, Python v. 3.7.9[17] was used. An overview of the NIfTI conversion workflow for segmentations and images is shown in Fig. 3. In brief, using an in-house Python script, .tag files (sliceOmatic output) were read in binary format and converted into numpy format[18], trimmed to remove header information, and then re-sized to the corresponding size of the 2D DICOM axial slice (sliceOmatic output) which was also converted to numpy format, i.e., a 2D array. The slice location was determined from the 2D DICOM axial slice in tandem with the 3D DICOM image (acquired from the TCIA) using pydicom[19]; the 3D DICOM image was necessary to determine the relative position of the 2D axial slice on the 3D volume. A 3D array that contained the segmentation information was then created by filling in all non-segmented slices with 0s, yielding a 3D segmentation mask. Each 3D segmentation mask contained separate regions of interest (0 = background, 1 = muscle, 2 = adipose for example in Fig. 3). A 3D representation was selected for the segmentation masks so that segmentations could be used for 3D applications (e.g., in tandem with the original 3D images), in addition to 2D applications (e.g., in tandem with single slice 2D images). 3D segmentation masks were converted to binary masks in NIfTI format (separate binary files for muscle and adipose) using SimpleITK[20]; separate binary files for each tissue type were generated for ease of use, e.g., most auto-segmentation approaches utilize binary masks[21]. 3D CT DICOM images were loaded into Python using the DICOMRTTool[22] library, and then converted to

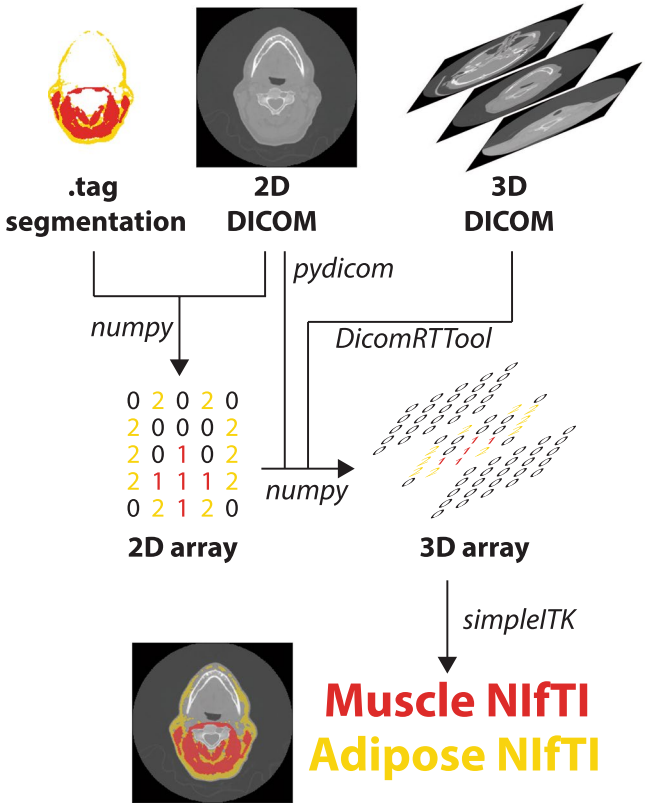


**Fig. 3** File conversion workflow for segmentations and images. Outputs from sliceOmatic software, i.e., .tag segmentation and 2D Digital Imaging and Communications in Medicine (DICOM) slice, are used to generate a 2D mask array of muscle and adipose tissue. Information from 2D DICOM slice and corresponding 3D DICOM image (acquired from corresponding The Cancer Imaging Archive dataset) are used to generate a 3D array, which is then converted to Neuroimaging Informatics Technology Initiative (NIfTI) format.

NIfTI format using SimpleITK. Additional documentation on scripts used for conversion can be located on the corresponding GitHub repository: https://github.com/kwahid/C3_sarcopenia_data_descriptor.

Of the 396 cases converted through the previously mentioned workflow, one patient (TCIA ID 0435) had a DICOM CT file with image reconstruction errors, while another (TCIA ID 0464) was unable to be converted to NIfTI format successfully, thus necessitating their removal from the final dataset, yielding 394 image/segmentation pairs in NIfTI format. Also worthy of note, 4 cases (TCIA ID's: 0226, 0280, 0577, and 0607) yielded partitioned segmentation masks (mask spread over several slices) secondary to export issues in sliceOmatic when loading images with oblique image orientations; these cases have been kept in the dataset for completeness but should likely not be used for most segmentation-related applications.

**Additional patient demographic data collection.** In addition to cross-sectional area derived from skeletal muscle segmentations, calculation of skeletal muscle index requires data concerning patient height and weight. In order to increase the usability of segmented regions of interest for sarcopenia-related calculations and model building, we also collected corresponding height (in m) and weight (in kg) data for all patients in our dataset. Anonymized TCIA IDs were mapped to existing patient medical record numbers to collect the corresponding data. Data were collected from the University of Texas MD Anderson Cancer Center clinical databases through the EPIC electronic medical record system by a manual review of clinical notes and paperwork. The Institutional Review Board of the University of Texas MD Anderson Cancer Center gave ethical approval for this work (RCR03–0800, waiver of informed consent). Height and weight were collected for the pre-radiotherapy visit only in accordance with the pre-radiotherapy imaging collected for this study. Clinical data collection was performed by a trained physician (D.E.).

## Data Records

**Segmentation data.** This data collection consists of 788 3D volumetric compressed NIfTI files (394 skeletal muscle "muscle.nii.gz" files, 394 adipose tissue "fat.nii.gz" files) derived from an original collection of 394 DICOM files of pre-therapy CT images collected from 495 TCIA cases ("Radiomics outcome prediction in Oropharyngeal cancer")[13,14]. The skeletal muscle and adipose tissue NIfTI files are binary masks (0 = background, 1 = tissue region of interest). While we do not provide the corresponding 394 CT images in NIfTI format due to Figshare upload size constraints, we do provide all the code necessary to produce these files (see Code availability section). In addition to NIfTI format files, we also include .tag segmentation files and corresponding 2D DICOM

files (sliceOmatic outputs) for interested parties to recreate our NIfTI conversion pipeline if desired. Of note, we do not include the 3D DICOM CT files as these can be acquired from existing TCIA repositories[13,14].

**Clinical data.** We also provide a single comma-separated value (CSV) file containing additional clinical demographic data germane to sarcopenia clinical-decision making. Within the CSV file, in addition to newly collected height and weight variables, we also include previously publicly available clinical variables in the TCIA dataset[13,14] relevant for body composition analysis (age and sex).

Segmentations are organized by an anonymized TCIA patient ID number ("TCIA Radiomics ID") and can be cross-referenced against the CSV data table using this identifier. The raw data, data records, and supplemental descriptions of the meta-data files are cited under Figshare: https://doi.org/10.6084/m9.figshare.18480917[23].

## Technical Validation

**Skeletal muscle segmentations.** The segmentations provided in this data descriptor have been utilized as ground-truth segmentations in a previous study by Naser *et al.*[12] which yielded sarcopenia determination results (normal vs. depleted skeletal muscle) that were consistent with existing literature[9], i.e., overall survival stratification is significant in males but not females as determined by Kaplan Meier analysis. Note: 4 patients included in the current data descriptor were excluded from the aforementioned analysis (TCIA ID's: 0226, 0280, 0577, and 0607), due to oblique image orientation mask issues previously described in Methods.

**EPIC (Electronic Medical Record System).** The University of Texas MD Anderson Cancer Center adopted this system in the year 2017 which allows integrating research data and accessing data from virtually every electronic source within the institution. https://www.clinfowiki.org/wiki/index.php/Epic_Systems.

## Usage Notes

This data collection is provided in NIfTI format with the accompanying CSV file containing additional clinical information indexed by TCIA identifier. We invite all interested researchers to download this dataset to use in sarcopenia-related research and automated clinical decision support tool development.

Images (reproducible through code) and segmentations are stored in NIfTI format and 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[24] and 3D Slicer[25].

## Code availability

Segmentation was performed using the commercially-available tool sliceOmatic v. 5.0 (Tomovision, Magog, Canada). The code for NIfTI file conversion of DICOM CT images and corresponding .tag format muscle/adipose tissue segmentations was developed using in-house Python scripts and is made publicly available through GitHub: https://github.com/kwahid/C3_sarcopenia_data_descriptor. Alternative code for converting .tag files to Matlab readable format can be located at: https://github.com/RJain12/matlab-tag-reader.

Received: 10 February 2022; Accepted: 22 July 2022;
Published online: 02 August 2022

## References

1. World Health Organization. *Global cancer observatory. International agency for research on cancer*. (2020).
2. van Rijn-Dekker, M. I. *et al*. Impact of sarcopenia on survival and late toxicity in head and neck cancer patients treated with radiotherapy. *Radiother. Oncol.* **147**, 103–110 (2020).
3. Findlay, M., White, K., Stapleton, N. & Bauer, J. Is sarcopenia a predictor of prognosis for patients undergoing radiotherapy for head and neck cancer? A meta-analysis. *Clin. Nutr.* **40**, 1711–1718 (2021).
4. Hua, X. *et al*. When the loss costs too much: a systematic review and meta-analysis of sarcopenia in head and neck cancer. *Front. Oncol.* **9**, 1561 (2020).
5. Perthen, J. E. *et al*. Intra-and interobserver variability in skeletal muscle measurements using computed tomography images. *Eur. J. Radiol.* **109**, 142–146 (2018).
6. Clark, K. *et al*. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
7. Kalpathy-Cramer, J., Freymann, J. B., Kirby, J. S., Kinahan, P. E. & Prior, F. W. Quantitative imaging network: data sharing and competitive AlgorithmValidation leveraging the cancer imaging archive. *Transl. Oncol.* **7**, 147–152 (2014).
8. Grossberg, A. J. *et al*. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Sci. Data* **5**, 1–10 (2018).
9. Olson, B. *et al*. Establishment and Validation of Pre-Therapy Cervical Vertebrae Muscle Quantification as a Prognostic Marker of Sarcopenia in Patients With Head and Neck Cancer. *Front. Oncol.* **12**, 812159 (2022).
10. Gregoire, V. *et al*. Intensity-modulated radiation therapy for head and neck carcinoma. (2007).
11. Glastonbury, C. M. Critical Changes in the Staging of Head and Neck Cancer. *Radiol. Imaging Cancer* **2**, e190022 (2020).
12. Naser, M. A. *et al*. Deep learning auto-segmentation of cervical skeletal muscle for sarcopenia analysis in patients with head and neck cancer. *Front. Oncol.* **12**, 930432 (2022).
13. Elhalawani, H. *et al*. Radiomics outcome prediction in Oropharyngeal cancer. *The Cancer Imaging Archive*. https://doi.org/10.7937/TCIA.2020.2vx6-fy46 (2018).
14. Elhalawani, H. *et al*. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci. Data* **4**, 170077 (2017).
15. Grossberg, A. J. *et al*. Association of body composition with survival and locoregional control of radiotherapy-treated head and neck squamous cell carcinoma. *JAMA Oncol.* **2**, 782–789 (2016).
16. 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).
17. Van Rossum, G. & Drake Jr, F. L. *Python reference manual*. (Centrum voor Wiskunde en Informatica Amsterdam, 1995).
18. Oliphant, T. E. *A guide to NumPy*. vol. 1 (Trelgol Publishing USA, 2006).

19. Mason, D. Pydicom: an open source DICOM library. *Med. Phys.* **38**, 3493–3493 (2011).
20. Lowekamp, B. C., Chen, D. T., Ibáñez, L. & Blezek, D. The design of SimpleITK. *Front. Neuroinformatics* **7**, 45 (2013).
21. Gros, C., Lemay, A. & Cohen-Adad, J. SoftSeg: Advantages of soft versus binary training for image segmentation. *Med. Image Anal.* **71**, 102038 (2021).
22. 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).
23. Wahid, K. *et al*. Skeletal muscle and adipose tissue at the C3 vertebral level for TCIA HNSCC patients. *Figshare*. https://doi.org/10.6084/m9.figshare.18480917.v1 (2022).
24. Abràmoff, M. D., Magalhães, P. J. & Ram, S. J. Image processing with ImageJ. *Biophotonics Int.* **11**, 36–42 (2004).
25. Fedorov, A. *et al*. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn. Reson. Imaging* **30**, 1323–1341 (2012).

## Acknowledgements

This work was supported by the National Institutes of Health (NIH)/National Cancer Institute (NCI) through a Cancer Center Support Grant (CCSG; P30CA016672–44). M.A.N. is supported by an NIH grant (R01DE028290–01). 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). A.J.G. received funding from the National Cancer Institute (K08 245188, R01 CA264133) and the American Association for Cancer Research/Mark Foundation "Science of the Patient" Award (20–60–51-MARK). B.O. received funding from the National Cancer Institute (F30 CA254033) and Radiologic Society of North America Research Medical Student Grant (RMS2026). V.S. received funding from The University of Texas, Graduate School of Biomedical Sciences Graduate research assistantship. 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).

## Author contributions

Study conceptualization: K.A.W., A.J.G., B.O., R.J., A.S.R.M., K.K., C.D.F., and M.A.N.; Study design: K.A.W., A.J.G., B.O., R.J., and A.S.R.M.; Data acquisition: A.J.G., B.O., R.J., D.E.-H., C.D., V.S., and M.A.; Quality control of data and algorithms: M.A.N., K.A.W., R.H., J.J., J.S., and K.K.; Manuscript editing: K.A.W., A.J.G., B.O., R.J., A.S.R.M., K.K., and M.A.N. All authors contributed to the article and approved the submitted version.

## Competing interests

C.D.F. has received direct industry grant support, speaking honoraria, and travel funding from Elekta AB. The other authors have no conflicts of interest to disclose.

## Additional information

**Correspondence** and requests for materials should be addressed to C.D.F. or M.A.N.

**Reprints and permissions information** is available at www.nature.com/reprints.

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