scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Annotated test-retest dataset of lung cancer CT scan images reconstructed at multiple imaging parameters

Binsheng Zhao[1]✉, Laurent Dercle[1,2], Hao Yang[1], Gregory J. Riely[1], Mark G. Kris[1] & Lawrence H. Schwartz[1]

Quantitative imaging biomarkers (QIB) are increasingly used in clinical research to advance precision medicine approaches in oncology. Computed tomography (CT) is a modality of choice for cancer diagnosis, prognosis, and response assessment due to its reliability and global accessibility. Here, we contribute to the cancer imaging community through The Cancer Imaging Archive (TCIA) by providing investigator-initiated, same-day repeat CT scan images of 32 non–small cell lung cancer (NSCLC) patients, along with radiologist-annotated lesion contours as a reference standard. Each scan was reconstructed into 6 image settings using various combinations of three slice thicknesses (1.25 mm, 2.5 mm, 5 mm) and two reconstruction kernels (lung, standard; GE CT equipment), which spans a wide range of CT imaging reconstruction parameters commonly used in lung cancer clinical practice and clinical trials. This holds considerable value for advancing the development of robust Radiomics, Artificial Intelligence (AI) and machine learning (ML) methods.

Background & Summary

Recent advances in artificial intelligence (AI) techniques such as radiomics, deep learning and machine-learning have turned standard of care CT images into a huge amount of numerical data that can be mined to extract imaging biomarkers. AI has the potential to measure and track every aspect of the information contained in CT images, regardless of visibility, and assess how they change over time. In this new era, AI can quantitatively characterize tumor imaging phenotypes (e.g., shape, heterogeneity, and vascularity) non-invasively and assess the inner organization processes of the entire tumor volume with the surrounding tissue.

The current literature has proven the concept that radiomics-guided precision medicine could improve the management of non–small cell lung cancer (NSCLC) patients. Virtual biopsies with radiomics have been envisioned to tackle the inherent risks and limitations of conventional invasive biopsies such as limited sampling site and significant false negative rate. Radiomics[1,2] has proven its worth for a wide range of indications including but not limited to diagnosis[3–6], prediction of histopathology[7–9], risk-stratification[10–13], prediction of outcome[14,15], prediction of mutational status[16–21], and response assessment to systemic cancer therapies[22–24].

The proof that AI interpretation of CT images can be used for precision medicine has created an unmet need to identify potential cofounding biases. To have any clinical utility and be qualified as an imaging biomarker, imaging features must provide clinical benefits to the patient, be sensitive and/or specific, and be robust and reproducible across multiple imaging acquisition platforms[25,26]. Pivotal pioneering efforts[27–32] identified several variables that influenced reproducibility in imaging features, which have been confirmed and further extended by recent reports[33–41]. These variables include CT acquisition parameters[16,32,42–44] (e.g., slice thickness, convolution kernel), segmentation tools of tumor lesions[45,46], and quality of contrast-enhancement[47–49]. Certain variables, however, have been identified to have high intra-patient reproducibility and inter-patient biological range[28] by test-retest and correlation analyses after a short-interval repeat scan.

In 2007, to assess the variability of tumor size measurements, we conducted an investigator-initiated same-day repeat CT study (a.k.a. coffee break study)[27]. In the study, 32 NSCLC patients underwent two CT scans

[1]Memorial Sloan-Kettering Cancer Center, New York, NY, 10021, USA. [2]Department of Radiology, Columbia University New York, New York, NY, 10032, USA. ✉e-mail: Zhaob1@mskcc.org

performed within 15 minutes. Each scan's raw data were reconstructed into six image series (one series corresponding to one image setting) using combinations of three slice thicknesses (1.25 mm, 2.5 mm, 5 mm) and two reconstruction kernels (lung, standard). This dataset will be referred as **M-setting Repeat Lung CT collection**. Using the **M-setting Repeat Lung CT** dataset, our team has published two pilot studies investigating the reproducibility of radiomic features affected by varying slice thicknesses and reconstruction kernels[30,32]. Leveraging publicly available feature values extracted by our team's feature extractor from the 6-setting repeat CT dataset[32], another research group has published a pioneering feature harmonization paper. This paper employs Combat technology to reduce variability induced by imaging parameters[50].

Among these 6 settings, one image setting reconstructed using 1.25 mm slice thickness and lung kernel was made publicly available in 2011 through the National Cancer Institute's (NCI) Reference Image Database to Evaluate Therapy Response (RIDER) project. Since then, this dataset, a.k.a. the **RIDER Lung CT collection**, has become a unique and frequently-downloaded dataset to evaluate computer-aided segmentation methods and explore the reproducibility and variability of quantitative imaging features (a.k.a. radiomic features), which resulted in numerous publications[2,27,28,51–53]. Given varying CT imaging protocols used in clinical practice and clinical trials for lung cancer, there is an urgent need to release our entire M-setting Repeat Lung CT collection, not only for retrospective analyses of the historical imaging data, but also for guidance of prospective radiomics studies in lung cancer. In addition to the repeat CT scan images from 32 NSCLC patients reconstructed at multiple settings, we are now also releasing annotated lesion contour images from one radiologist as a reference.

Methods

Patient recruitment. The Institutional Review Board of Memorial Sloan-Kettering Cancer Center (MSKCC), where the original clinical trial (ClinicalTrials.gov identifier NCT00579852) was conducted and the data was collected, approved this Health Insurance Portability and Accountability Act compliant study (IRB#: 06-149). Informed consent for participation in this study and potential publication of study outcomes was obtained from all patients[27]. In 2006, when this study was conducted, sharing medical images was not a common practice. However, our IRB leadership has recently granted approval for the sharing of deidentified repeat CT data.

From January 2007 to September 2007, we recruited a total of 32 consecutive qualified adult patients with lung cancer who were under the care of oncologists. Patients were deemed eligible for the trial if they had pathologically confirmed NSCLC with primary pulmonary tumors of 1 cm or larger and were determined to need unenhanced chest CT in the near future. The demographics of the 32 patients were: mean age of 62.1 years (range: 29–82 years), including 16 men with mean age of 61.8 years (range: 29–79 years) and 16 women with mean age of 62.4 years (range: 45–82 years).

Repeat CT image data. Each patient received the scheduled chest CT scan and then was asked to leave the scanner table to walk around the CT scanner site. Afterwards, they returned for a second scan of the chest using the same imaging protocol. The two scans of the patients were performed within 15 minutes on the same scanner, either LightSpeed 16 (16-row) or VCT (64-row) (GE Medical Systems, Milwaukee, WIS).

The standard-dose thoracic images were obtained without intravenous contrast while holding breath. Parameters of the 16-detector row scanner (or 64-detector row scanner) were set as the followings: detector configuration, 16×1.25 mm (64×0.63 mm); tube voltage, 120 kVp; tube current, 299–441 mA (298–351 mA); pitch of 1.375:1 (0.984:1). Each of the two scans was reconstructed using one of the three slice thicknesses of 1.25 mm, 2.5 mm, or 5 mm and one of the reconstruction kernels of lung or standard.

Lesion segmentation. Thirty-two lung cancer lesions, one from each of the 32 patients, were segmented by an experienced radiologist with the help of our in-house lung lesion segmentation algorithm[54] that had been integrated into an image analysis platform built upon an open source Weasis[55]. The readings were split into 12 sessions; one session for one image setting (2 repeat scans $\times$ 6 image settings/scan). The time interval between any two reading sessions was between 2 and 3 weeks to reduce the memory bias. Using the contour editing tool provided by the image analysis platform, the radiologist was allowed to correct computer-generated results/contours when they were suboptimal. Figure 1 illustrates an example of a lung tumor as captured in a CT scan, reconstructed at 6 different image settings (Fig. 1 upper panel (a-f)) alongside their corresponding segmentations (Fig. 1 lower panel (a1-f1)). More details on the six-setting repeat CT scan dataset and the lesion segmentation can be found in the references[30].

De-Identification of imaging data. The repeat CT scan images are in DICOM format and lesion annotated images are in DICOM SEG format. All images are de-identified using DICOM Supplement 142:Clinical Trial de-identification Profiles.

Data Records

The dataset is available at The Cancer Imaging Archive[56].

Subject identifiers. A unique identifier for each subject is identical in all our datasets and in the form of RIDER-XXX. And all the subjects have the same "*RIDER LUNG CT*" collection ID.

RIDER Lung CT (one-setting repeat CT images). One of the six image settings of the repeat CT scan images have been already stored in TCIA since 2011. The CT images are in DICOM format. There is an accompanying note file in Microsoft Excel XLS format, containing lesion RIDER ID, example image, and coordinates in both (repeat) scan images (Fig. 2).

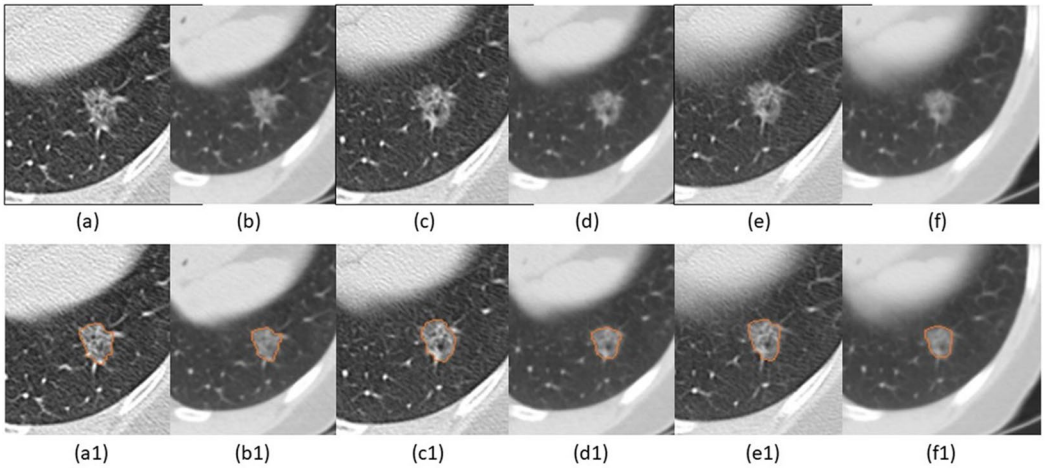

Fig. 1 A lung tumor from a CT scan, reconstructed with six image settings, presented alongside their segmentations. The upper panel shows cropped original CT images containing a tumor, while the lower panel shows the same images with their corresponding contours superimposed. 1.25 mm slice thickness with the lung reconstruction algorithm (sharp image) (1.25 L) (**a**; a1) and the standard reconstruction algorithm (smooth image) (1.25S) (**b**; b1); 2.5 mm slice thickness with lung reconstruction (2.5 L) (**c**; c1) and standard reconstruction (2.5S) (**d**; d1); 5 mm slice thickness with lung reconstruction (5 L) (**e**; e1) and standard reconstruction (5S) (**f**; f1).

MSKCC Coffe Break study conducted in 2007

Case#		Scan #1 Location of a voxel around the lesion center (origin: top-left)					Scan #2 (repeat scan) Location of a voxel around the lesion center (origin: top-left)			
	RIDER-ID	x (in pixel)	y(in pixel)	z (imagenumber)	z_imageposition(mm)	Example image	x (in pixel)	y(in pixel)	z (imagenumber)	z_imageposition(mm)
1	RIDER-1500037140	135	191	47	-62.5		143	180	47	-72.5

Fig. 2 Microsoft Excel XLS file containing the lesion notes.

CT scans	Six settings of three slice thicknesses and two reconstruction kernels					
	1.25L	1.25S	2.5L	2.5S	5L	5S
First CT scan	+	+	+	+	+	+
Second CT scan	+	+	+	+	+	+
RIDER Lung CT collection	+	—	—	—	—	—
M-setting Repeat Lung CT collection	+	+	+	+	+	+

Table 1. legend. Reconstructed six image settings of each scan. The six settings are 1.25L, 1.25S, 2.5L, 2.5S, 5L, and 5S, where 1.25, 2.5 and 5 are the three slice thicknesses and L and S indicate the lung and the standard reconstruction kernels, respectively. The "+" and "—" signs indicate whether a dataset includes or does not include the specific setting above.

M-setting repeat lung CT. The six settings of the repeat CT scan images have been stored in TCIA since June, 2024 (Table 1). CT images are in DICOM format. Lesion annotation images are in DICOM SEG format. The same accompanying note file will help to identify the lesions in the patients.

Image data. All the CT images of this *M-setting Repeat Lung CT* collection are de-identified and in DICOM format.

Technical Validation

All CT image data were collected as part of standard patient care. Image quality assurance was performed regularly at the institution where the data were collected. One of the six-setting repeat CT scan images, the RIDER Lung CT collection, is already stored in TCIA and widely downloaded and used by people all over the world.

Usage Notes

Currently, the RIDER Lung CT collection is a unique publicly available same-day repeat CT image dataset serving to study the reproducibility of radiomics features. This six-setting repeat CT lung cancer image dataset will allow more comprehensive investigation of the reproducibility and robustness across imaging acquisition parameters for AI techniques such as radiomics, deep learning, and machine-learning methods. It also provides a valuable dataset to help identify preferred CT imaging parameter settings for better studying AI/radiomics in lung cancer.

Code availability

No custom code has been used.

Received: 25 March 2024; Accepted: 5 November 2024;
Published online: 20 November 2024

References

1. Lambin, P. *et al*. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer* **48**, 441–446 (2012).
2. Aerts, H. J. *et al*. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **5**, 1–9 (2014).
3. Hawkins, S. *et al*. Predicting Malignant Nodules from Screening CT Scans. *J Thorac Oncol* **11**, 2120–2128, https://doi.org/10.1016/j.jtho.2016.07.002 (2016).
4. Liu, Y. *et al*. Radiological Image Traits Predictive of Cancer Status in Pulmonary Nodules. *Clin Cancer Res* https://doi.org/10.1158/1078-0432.CCR-15-3102 (2016).
5. Setio, A. A. *et al*. Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks. *IEEE Trans Med Imaging* **35**, 1160–1169, https://doi.org/10.1109/TMI.2016.2536809 (2016).
6. Shen, W. *et al*. Multi-crop convolutional neural networks for lung nodule malignancy suspiciousness classification. *Pattern Recognition* **61**, 663–673 (2017).
7. Tavaré, R. *et al*. An effective immuno-PET imaging method to monitor CD8-dependent responses to immunotherapy. *Cancer research* **76**, 73–82 (2016).
8. Coroller, T. P. *et al*. Radiomic phenotype features predict pathological response in non-small cell lung cancer. *Radiother Oncol* **119**, 480–486, https://doi.org/10.1016/j.radonc.2016.04.004 (2016).
9. Lu, L. *et al*. A quantitative imaging biomarker for predicting disease-free-survival-associated histologic subgroups in lung adenocarcinoma. *European Radiology*, 1–10 (2020).
10. Fried, D. V. *et al*. Prognostic value and reproducibility of pretreatment CT texture features in stage III non-small cell lung cancer. *Int J Radiat Oncol Biol Phys* **90**, 834–842, https://doi.org/10.1016/j.ijrobp.2014.07.020 (2014).
11. Coroller, T. P. *et al*. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. *Radiother Oncol* **114**, 345–350, https://doi.org/10.1016/j.radonc.2015.02.015 (2015).
12. Ganeshan, B., Panayiotou, E., Burnand, K., Dizdarevic, S. & Miles, K. Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival. *European radiology* **22**, 796–802 (2012).
13. Yoon, H. J. *et al*. Decoding tumor phenotypes for ALK, ROS1, and RET fusions in lung adenocarcinoma using a radiomics approach. *Medicine* **94** (2015).
14. Huang, Y. *et al*. Radiomics Signature: A Potential Biomarker for the Prediction of Disease-Free Survival in Early-Stage (I or II) Non-Small Cell Lung Cancer. *Radiology* **281**, 947–957, https://doi.org/10.1148/radiol.2016152234 (2016).
15. Emaminejad, N. *et al*. Fusion of Quantitative Image and Genomic Biomarkers to Improve Prognosis Assessment of Early Stage Lung Cancer Patients. *IEEE Trans Biomed Eng* https://doi.org/10.1109/TBME.2015.2477688 (2015).
16. Li, Y. *et al*. CT slice thickness and convolution kernel affect performance of a radiomic model for predicting EGFR status in non-small cell lung cancer: a preliminary study. *Scientific reports* **8**, 1–10 (2018).
17. Liu, Y. *et al*. Radiomic features are associated with EGFR mutation status in lung adenocarcinomas. *Clinical lung cancer* **17**, 441–448. e446 (2016).
18. Kim, T. J., Lee, C.-T., Jheon, S. H., Park, J.-S. & Chung, J.-H. Radiologic characteristics of surgically resected non-small cell lung cancer with ALK rearrangement or EGFR mutations. *The Annals of thoracic surgery* **101**, 473–480 (2016).
19. Zhou, J. *et al*. Comparative analysis of clinicoradiologic characteristics of lung adenocarcinomas with ALK rearrangements or EGFR mutations. *European radiology* **25**, 1257–1266 (2015).
20. Yang, Y. *et al*. EGFR L858R mutation is associated with lung adenocarcinoma patients with dominant ground-glass opacity. *Lung cancer* **87**, 272–277 (2015).
21. Hsu, J.-S. *et al*. Correlation between EGFR mutation status and computed tomography features in patients with advanced pulmonary adenocarcinoma. *Journal of thoracic imaging* **29**, 357–363 (2014).
22. Aerts, H. J. *et al*. Defining a Radiomic Response Phenotype: A Pilot Study using targeted therapy in NSCLC. *Sci Rep* **6**, 33860, https://doi.org/10.1038/srep33860 (2016).
23. Dercle, L. *et al*. Identification of Non-Small Cell Lung Cancer Sensitive to Systemic Cancer Therapies Using Radiomics. *Clin Cancer Res* https://doi.org/10.1158/1078-0432.CCR-19-2942 (2020).
24. Shi, L. *et al*. Radiomics for response and outcome assessment for non-small cell lung cancer. *Technology in cancer research & treatment* **17**, 1533033818782788 (2018).
25. Huang, E. P. *et al*. Criteria for the translation of radiomics into clinically useful tests. *Nat Rev Clin Oncol* **20**, 69–82, https://doi.org/10.1038/s41571-022-00707-0 (2023).
26. Zhao, B. Understanding Sources of Variation to Improve the Reproducibility of Radiomics. *Front Oncol* **11**, 633176, https://doi.org/10.3389/fonc.2021.633176 (2021).
27. Zhao, B. *et al*. Evaluating variability in tumor measurements from same-day repeat CT scans of patients with non-small cell lung cancer. *Radiology* **252**, 263–272, https://doi.org/10.1148/radiol.2522081593 (2009).
28. Balagurunathan, Y. *et al*. Test–retest reproducibility analysis of lung CT image features. *Journal of digital imaging* **27**, 805–823 (2014).
29. Zhao, B., Tan, Y., Tsai, W. Y., Schwartz, L. H. & Lu, L. Exploring variability in CT characterization of tumors: a preliminary phantom study. *Translational oncology* **7**, 88 (2014).
30. Zhao, B. *et al*. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific reports* **6**, 1–7 (2016).
31. Mackin, D. *et al*. Measuring CT scanner variability of radiomics features. *Investigative radiology* **50**, 757 (2015).
32. Lu, L., Ehmke, R. C., Schwartz, L. H. & Zhao, B. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PloS one* **11** (2016).
33. Welch, M. L. *et al*. Vulnerabilities of radiomic signature development: the need for safeguards. *Radiotherapy and Oncology* **130**, 2–9 (2019).

34. Lu, L., Liang, Y., Schwartz, L. H. & Zhao, B. Reliability of Radiomic Features Across Multiple Abdominal CT Image Acquisition Settings: A Pilot Study Using ACR CT Phantom. *Tomography* **5**, 226 (2019).
35. Kim, H. *et al*. Effect of reconstruction parameters on the quantitative analysis of chest computed tomography. *Journal of thoracic imaging* **34**, 92–102 (2019).
36. Huang, Y. *et al*. Radiomics signature: a potential biomarker for the prediction of disease-free survival in early-stage (I or II) non—small cell lung cancer. *Radiology* **281**, 947–957 (2016).
37. Talwar, A. *et al*. Pulmonary nodules: Assessing the imaging biomarkers of malignancy in a "coffee-break. *European journal of radiology* **101**, 82–86 (2018).
38. Berenguer, R. *et al*. Radiomics of CT features may be nonreproducible and redundant: influence of CT acquisition parameters. *Radiology* **288**, 407–415 (2018).
39. Mackin, D. *et al*. Effect of tube current on computed tomography radiomic features. *Scientific reports* **8**, 1–10 (2018).
40. Haga, A. *et al*. Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. *Radiological physics and technology* **11**, 27–35 (2018).
41. Shafiq-ul-Hassan, M. *et al*. Accounting for reconstruction kernel-induced variability in CT radiomic features using noise power spectra. *Journal of Medical Imaging* **5**, 011013 (2017).
42. He, L. *et al*. Effects of contrast-enhancement, reconstruction slice thickness and convolution kernel on the diagnostic performance of radiomics signature in solitary pulmonary nodule. *Scientific reports* **6**, 34921 (2016).
43. Dercle, L. *et al*. Limits of radiomic-based entropy as a surrogate of tumor heterogeneity: ROI-area, acquisition protocol and tissue site exert substantial influence. *Scientific reports* **7**, 1–10 (2017).
44. Dercle, L. *et al*. Radiomics Response Signature for Identification of Metastatic Colorectal Cancer Sensitive to Therapies Targeting EGFR Pathway. *JNCI: Journal of the National Cancer Institute* (2020).
45. Parmar, C. *et al*. Robust radiomics feature quantification using semiautomatic volumetric segmentation. *PloS one* **9** (2014).
46. Huang, Q. *et al*. Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. *Journal of Medical Imaging* **5**, 011005 (2017).
47. Dercle, L. *et al*. Impact of variability in portal venous phase acquisition timing in tumor density measurement and treatment response assessment: metastatic colorectal cancer as a paradigm. *JCO clinical cancer informatics* **1**, 1–8 (2017).
48. Ma, J. *et al*. Automated Identification of Optimal Portal Venous Phase Timing with Convolutional Neural Networks. *Academic radiology* **27**, e10–e18 (2020).
49. Dercle, L. *et al*. Using a single abdominal Computed Tomography image to differentiate five contrast-enhancement phases: a machine-learning algorithm for radiomics-based precision medicine. *European Journal of Radiology*, 108850 (2020).
50. Orlhac, F., Frouin, F., Nioche, C., Ayache, N. & Buvat, I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology* **291**, 53–59, https://doi.org/10.1148/radiol.2019182023 (2019).
51. Oxnard, G. R. *et al*. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. *Journal of Clinical Oncology* **29**, 3114 (2011).
52. McNitt-Gray, M. F. *et al*. Determining the variability of lesion size measurements from CT patient data sets acquired under "no change" conditions. *Translational oncology* **8**, 55–64 (2015).
53. Buckler, A. J. *et al*. Inter-method performance study of tumor volumetry assessment on computed tomography test-retest data. *Academic radiology* **22**, 1393–1408 (2015).
54. Tan, Y., Schwartz, L. H. & Zhao, B. Segmentation of lung lesions on CT scans using watershed, active contours, and Markov random field. *Med Phys* **40**, 043502, https://doi.org/10.1118/1.4793409 (2013).
55. Yang, H., Schwartz, L. H. & Zhao, B. A Response Assessment Platform for Development and Validation of Imaging Biomarkers in Oncology. *Tomography* **2**, 406–410, https://doi.org/10.18383/j.tom.2016.00223 (2016).
56. Zhao, B., Schwartz, L. H., Kris, M. G. & Riely, G. J. Coffee-break lung CT collection with scan images reconstructed at multiple imaging parameters (Version 3) [Dataset]. *The Cancer Imaging Archive*. https://doi.org/10.7937/k9/tcia.2015.u1x8a5nr (2024).

Author contributions

Data acquisition: B.Z., H.Y., G.J.R., M.G.K., L.H.S.; Manuscript drafting and figures/tables preparing: L.D., B.Z.; Manuscript revision and/or approval: all authors.

Competing interests

B.Z.: Patents, Royalties: Varian Medical Systems; L. H. S.: Consulting or Advisory Role: Novartis, GlaxoSmithKline. Research Funding: Eli Lilly (Inst), Astellas Pharma (Inst), Merck Sharp & Dohme (Inst), Pfizer (Inst). Patents, Royalties: Varian Medical Systems. All authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.Z.

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.