Identification of Precise 3D CT Radiomics for Habitat Computation by Machine Learning in Cancer

Olivia Prior, MSc[1], Carlos Macarro, MSc[1], Víctor Navarro, BSc[2], Camilo Monreal, MSc[1], Marta Ligero, MSc[1], Alonso Garcia-Ruiz, MSc[1], Garazi Serna, PhD[2], Sara Simonetti, MD, PhD [2], Irene Braña, MD, PhD[1,2], Maria Vieito, MD, PhD[1,2], Manuel Escobar, MD[2], Jaume Capdevila, MD[2], Annette T. Byrne, MD, PhD[3,4], Rodrigo Dienstmann, MD[1], Rodrigo Toledo, PhD[1], Paolo Nuciforo, MD, PhD[1], Elena Garralda, MD, MSc[1,2], Francesco Grussu, PhD[1], Kinga Bernatowicz, PhD[1*], Raquel Perez-Lopez, MD, PhD[1,2*†]

[1] Vall d'Hebron Institute of Oncology (VHIO), 08035, Barcelona, Spain
[2] Vall d'Hebron University Hospital (VHUH), 08035, Barcelona, Spain
[3]Department of Physiology and Medical Physics, Centre for Systems Medicine, Royal College of Surgeons in Ireland, Dublin, Ireland
[4] National Pre-clinical Imaging Centre, Ireland

*KB and RPL are co-senior authors
[†]Corresponding author: *rperez@vhio.net, +34 689648377, Radiomics Group, Vall d'Hebron Institute of Oncology (VHIO), 08035, Barcelona, Spain

Funding: RPL is supported by LaCaixa Foundation, a CRIS Foundation Talent Award (TALENT19-05), the FERO Foundation, the Instituto de Salud Carlos III-Investigacion en Salud (PI18/01395 and PI21/01019), the Prostate Cancer Foundation (18YOUN19) and the Asociación Española Contra el Cancer (AECC) (PRYCO211023SERR, funding CM). ML is supported by PERIS PIF-Salut Grant. FG was funded by the Govt of Catalonia (Beatriu de Pinós 2020 BP 00117) and now receives the support of a fellowship from" la Caixa" Foundation (ID 100010434). The fellowship code is "LCF/BQ/PR22/11920010". KB was funded by the Govt of Catalonia (Beatriu de Pinós 2019 BP 00182). OPP was funded by Fundació "LaCaixa" (ID 100010434, code LCF/BQ/DR21/11880027). ATB, RD, PN and RPL are supported by funding from the EU Horizon 2020 Research and Innovation Programme under Grant Agreement number 754923 (COLOSSUS). This study received support from AstraZeneca (AZ). AZ was not involved in the study design; collection, analysis, interpretation of data; manuscript writing and decision to submit the manuscript for publication; or any other aspect concerning this work.

Manuscript Type: original research

Word Count for Text: 3097

Data sharing statement: The datasets used in this study are not publicly available since patients did not sign a consent form authorizing the public release of their data, even in anonymized form.

Conflicts of interest: RD declares advisory role for Roche, Boehringer Ingelheim, received a speaker's fee from Roche, Boehringer Ingelheim, Ipsen, Amgen, Servier, Sanofi, Libbs, Merck Sharp & Dohme, Lilly, AstraZeneca, Janssen, Takeda and research grants from Merck, Novartis, Daiichi-Sankyo, Pierre Fabre and AstraZeneca. All other authors have nothing to declare.

SUMMARY STATEMENT

Tumor heterogeneity was evaluated by computing stable CT tumor habitats with unsupervised learning on repeatable and reproducible 3D radiomics features in lung and liver cancer lesions.

KEY POINTS

- In this retrospective study of 2436 tumoral lesions from 605 CT scans, 3D radiomics features showed poor repeatability (median lower limit of the 95% confidence interval of the intraclass correlation coefficient, 0.442) and reproducibility against kernel radius (0.440) but excellent reproducibility against bin size (0.929).

- Out of 91 3D radiomics features analyzed, 26 were identified as precise (i.e., featuring moderate, good or excellent repeatability and reproducibility), with different sets of precise features identified in lung and liver lesions; habitats obtained with the subsets of precise radiomics features were more stable than those obtained with all computed features in both lung and liver lesions (13% and 11% increase in median Dice similarity coefficient in lung and liver lesions, P<.001).

- In an exploratory case study, CT habitats correlated well with multiparametric MRI habitats and histology, capturing intra-tumoral heterogeneity (e.g., areas with different tumor cell density and vascularization, or characterized by necrosis or fibrosis).

Abbreviations

ICC: Intraclass Correlation Coefficient
LCL: Lower 95% confidence limit of the Intraclass Correlation Coefficient
mpMRI: multiparametric MRI
DSC: Dice similarity coefficient
RF: Radiomics feature
HU: Hounsfield units
GMM: Gaussian mixture model

Identification of Precise 3D CT Radiomics for Habitat Computation by Machine Learning in Cancer

ABSTRACT

Purpose: To identify precise 3D radiomics features in CT images that enable computation of stable and biologically meaningful habitats with machine learning for cancer heterogeneity assessment.

Materials and Methods: This retrospective study included 2436 liver or lung lesions from 605 CT scans (November 2010-December 2021) in 318 patients with cancer (mean age, 64.5 years ± 10.1 [SD]; 185 male patients). 3D radiomics were computed from original and perturbed (simulated retest) images with different combinations of feature computation kernel radius (R) and bin size (B). The lower 95% confidence limit (LCL) of the intraclass correlation coefficient was used to measure repeatability and reproducibility. Precise features were identified by combining repeatability and reproducibility results (LCL≥0.50). Habitats were obtained with Gaussian Mixture Models in original and perturbed data using precise radiomics features and compared with habitats obtained using all features. The Dice similarity coefficient (DSC) was used to assess habitat stability. Biological correlates of CT habitats were explored in a case study, with a cohort of 13 patients with CT, multiparametric MRI (mpMRI) and tumor biopsies.

Results: 3D radiomics showed poor repeatability (median LCL[IQR] of 0.442 [0.312-0.516]) and poor reproducibility against R (0.44[0.33-0.526]) but excellent reproducibility against B (0.929[0.853-0.988]). Twenty-six radiomics features were precise, differing in lung and liver lesions. Habitats obtained with precise features (DSC, 0.532 (0.424-0.637 and 0.587 (0.465-0.703) for lung and liver lesions, respectively) were more stable than those obtained with all features (DSC, 0.601 (0.494-0.712) and 0.651 (0.52-0.784), respectively; *P* < .001). In the case study, CT habitats correlated quantitatively and qualitatively, with heterogeneity observed in mpMRI habitats and histology.

Conclusion: Precise 3D radiomics features were identified on CT that enabled tumor heterogeneity assessment through stable tumor habitat computation.

INTRODUCTION

Intra-tumoral spatial heterogeneity is a well-known characteristic of cancer (1). At the smallest scales, such heterogeneity refers to regions with diverse clones and tumor micro-environment harboring various levels of genomic and transcriptomic expressions. At macroscopic scales, this translates to tumor niches (i.e., subpopulation of cells localized in a particular intratumoral region) and tissue types (i.e., fibrosis, necrosis) (2). Such heterogeneity poses a challenge to targeted therapies, as distinct intra-tumoral regions may develop resistance to treatment (3,4). As a result, new lines of research have emerged to detect and quantify such heterogeneity non-invasively. A notable example, holding promise in the clinical setting, is CT-based habitat imaging, which aims to identify spatial regions (habitats) that exhibit shared imaging phenotypes in CT scans (5,6). The main advantage of CT tumor habitats is their ability to capture heterogeneity of the whole tumor (i.e., three-dimensionally) non-invasively. As such, in the age of precision medicine, CT tumor habitats could serve as a valuable tool to detect treatment-resistant habitats (and thus improve treatment selection), as well as to monitor response and tumor evaluation longitudinally and repeatedly throughout disease progression (7).

Successful implementation of tumor habitats in the clinic requires robust voxel-wise or 3D radiomics features (RFs) (8), which are the underlying imaging texture features used to generate them. In other words, RFs should be both repeatable (i.e., exhibiting measurement precision under the same set of computation conditions, also known as test-retest) and reproducible (i.e., exhibiting measurement precision under different computation conditions) (9,10). However, current literature on precision of CT radiomics is focused on RF as independent predictive/prognostic bi-omarkers and is limited regarding cancer types and reporting quality, with many studies neglecting to provide critical information such as whether texture RF were computed in 2D or 3D (11,12). This is especially relevant for habitat computation since 2D RF, which are computed disregarding neighboring voxels on out-of-plane slices, are less representative of tumor heterogeneity (13,14). Thus, there is a lack of knowledge on precision of 3D RF for CT tumor habitat computation.

In this study, we aimed to fill this knowledge gap and assess how 3D RF are affected by three different sources of variability: (i) test-retest scenarios, (ii) changes in kernel radius, R, which specifies the number of neighboring voxels to be considered when computing features, (iii) and changes in bin size, B, which defines the number of grey levels prior to feature computation. Features showing acceptable repeatability and reproducibility against the two computation parameters were identified as precise. We developed a Gaussian mixture model (GMM)-based unsupervised machine learning model (15,16) for tumor habitat computation and studied whether the use of precise RFs resulted in more stable habitats. Lastly, we explored the biological correlates of CT habitats in an independent cohort with multiparametric MRI (mpMRI) and digitized images of hematoxylin and eosin (HE)-stained biopsies. Our main goal was to develop a method to compute stable lung and liver tumor habitats based on the identified precise 3D radiomic features and GMMs to assess intratumoral heterogeneity.

MATERIALS AND METHODS

Patient Cohorts

We retrospectively analyzed 2436 lesions (1861 liver and 575 lung) from 318 patients (mean age, 64.5 years ± 10.1 [SD]; 185 male patients) with CT scans at multiple timepoints (605 total CT scans [Table 1]). Intravenous contrast-enhanced CT scans from patients with advanced cancer and lung or liver tumors from November 2010 to December 2021 were included. The analysis of these anonymized CT images was approved by the Vall d'Hebron Ethics Committee with waiver of informed consent. The study sample was split into four different cohorts (Figure 1A) depending on primary tumor location: 1) colorectal, 2) lung, 3) gastrointestinal neuroendocrine tumors and 4) a cohort including a mix of other cancers. Patients with gastrointestinal neuroendocrine tumors were selected from the multicenter phase II TALENT trial (NCT02678780). Details of patient cohorts and imaging protocols are reported in Table E1, Table E2, and Table E3 (online). Patients from the independent cohort with CT, mpMRI and biopsies, who were included in our case study, were included in the PREDICT prospective trial (also approved by the Vall d'Hebron Ethics Committee, PR(AG)29/2020) and signed consent for the acquisition and analysis of the CT and MRI scans and tumors biopsies. Details regarding imaging protocols and clinical information are reported in Appendix E1 and Tables E4 and E5 (online).

Image Segmentation, Perturbation and Feature Computation

An experienced radiologist (RPL) with more than 10 years of experience in oncological imaging segmented the entire volume of all measurable lesions according to the Response Evaluation Criteria version 1.1. (RECIST 1.1) (17) (i.e., maximal diameter ≥ 10mm) using 3D Slicer (v4.11.20210226, https://www.slicer.org/) (18). We assessed repeatability by simulating a retest scenario with image perturbation using the Medical Image Radiomics Processor (MIRP) Python toolkit (v1.2.0, https://github.com/oncoray/mirp) (19). Details are provided in Appendix E2. Original (i.e. no-filter) RF were computed for every lesion using PyRadiomics (v3.0.1, https://github.com/AIM-Harvard/pyradiomics/) (20). All RF were computed four times per lesion, each time using a different combination of settings for R (1mm/3mm) and B (12HU/25HU), hereinafter referred to as R1B12, R1B25, R3B12 and R3B25. We selected R and B values based on common practices, including those used by PyRadiomics, which defaults to R=1mm and B=25HU—these served as our primary reference points. Additionally, we selected B=12 Hounsfield units (HU) and R=3mm, as these values are also commonly employed in the field. In total, 91 features were analyzed (full list available in Table E4 [online]). Examples of RF are displayed in Figure E1, and relevant computation details are reported in Table E5 and Appendix E3 (online), in compliance with the Image Biomarker Standardization Inititative (IBSI) (21).

Repeatability and Reproducibility Analyses to Identify Precise Features

Figure 1B displays the precision analysis overview. We studied repeatability in four experiments, comparing radiomics values from original and perturbed (simulated retest) CT scans for each of the four setting combinations used (R1B12, R1B25, R3B12 and R3B25). For reproducibility against R, we conducted two experiments: first, we compared original RF computed with different R and fixed B=12HU (i.e., R1B12 vs R3B12); second, we compared original radiomics values computed with different R and fixed B=25HU (i.e., R1B25 vs R3B25). Similarly, we studied reproducibility against B twice: first for original RF computed with fixed R=1mm (R1B12 vs R1B25) and then for original radiomics values computed with fixed R=3mm (R3B12 vs R3B25). We conducted all experiments for all lesions and cohorts combined, and then for lung and liver lesions separately, as well as for each cohort (different primary tumor types) separately to understand the effect of primary tumor and lesion location on precision. The intraclass correlation coefficient (ICC) (22) was used to measure repeatability and reproducibility of features. A feature was selected as precise if the 95% lower confidence limit (LCL) of the ICC was ≥ 0.50 across the three relevant experiments: repeatability (R3B12), reproducibility against R (B=12HU), and reproducibility against B (R=3mm).

Imaging Habitats Computation

Habitats were computed for all lesions, four times per lesion, using either precise RF or non-precise RF (I.e., all computed RF), in both original and perturbed images. Prior to computation, Spearman's rank correlation coefficient (r) was used to eliminate highly correlated features ($r \geq 0.7$) (23) at a significance value of $p<.001$. Nonredundant radiomics were clustered with GMMs to obtain habitats. The optimal number of habitats was found using the Bayesian Information Criterion (BIC) (24). The stability of habitats (i.e., similarity of habitats computed in each original-perturbed pair) was measured according to the Dice Similarity Coefficient (DSC). See Appendix E4 (online) for more details.

As an exploratory case study, we evaluated the biological relevance of CT imaging habitats in an independent cohort of 13 patients with CT, mpMRI (including anatomical, diffusion-weighted and perfusion MRI), and digitized Hematoxylin-Eosin (HE) images from biopsy. Methods and results related to this case study are available in Appendix E5 (online).

Statistical Analysis

Following Koo's guidelines (22), we used the ICC based on a single-measurement, absolute-agreement t, two-way mixed-effects model for the repeatability analysis. For reproducibility, we computed the ICC based on a single-measurement, consistency, two-way mixed-effects model. We assigned each features' repeatability and reproducibility to the following categories based on the 95% LCL of the ICC: poor ($LCL < 0.5$); moderate ($0.5 \leq LCL< 0.75$); good ($0.75 \leq LCL < 0.9$); and excellent ($LCL \geq 0.90$) following (22) and (21). More details are provided in Appendix E6 (online). A paired two-sided Wilcoxon Signed-Rank test was conducted to evaluate the significance of differences in feature reproducibility between computation parameters, lesion location and habitat stability with precise features or all features. The effect size of the tests was calculated with Cohen's *d* and defined as small ($d \geq 0.20$), medium ($d \geq 0.50$), or large ($d \geq 0.80$) (25). A P value of $p<.05$ was used as the threshold for statistical significance.

All statistical tests were reviewed by VN (statistician) and performed using Python (v3.7.10; Python Software Foundation, Delaware, USA). All codes can be publicly accessed at https://github.com/radiomicsgroup/precise-habitats.

RESULTS

Repeatability Analysis

For every combination of computation settings (R1B12, R1B25, R3B12, and R3B25), the ICC was computed between the radiomics values computed from original (*test)* and perturbed (*simulated retest)* paired CT scans. Results showed that radiomics computed with R3B12 had the highest repeatability, with median (interquartile range) of the ICC 95% LCL for all features of 0.442 (0.312-0.516), compared with 0.191 (0.116-0.382), 0.199 (0.103-0.344) and 0.415 (0.306-0.516) for settings R1B12, R1B25 and R3B25, respectively. Figure 2A shows the proportions of RF with poor ($LCL < 0.5$), moderate ($0.5 \leq LCL < 0.75$), good ($0.75 \leq LCL < 0.9$), and excellent ($LCL \geq 0.90$) repeatability.

Regarding the effect of primary tumor and lesion location on repeatability, no evidence of differences in radiomics repeatability were found between different primary cancers for any of the settings (Figure E3). Moreover, while there were no major differences between liver and lung lesions in terms of the proportions of repeatable radiomics (Table E7[online]), the type of repeatable radiomics differed (Figure 2B). For instance, first-order and Gray-Level Run-Length Matrix (GLRLM) features were more repeatable in liver lesions than in lung, while Gray Level Co-occurrence Matrix (GLCM) features were more repeatable in lung lesions. Figure E4 (online) displays repeatability results for all features in the four repeatability experiments.

Reproducibility Analysis

Overall, radiomics values were more affected by changes in R than by changes in B (i.e., RF were more reproducible against B than against R). RF computed with a fixed B of 12HU were more reproducible than those computed with a fixed B of 25HU (Figure 3A), with median (interquartile range) of the LCL for all features of 0.440 (0.330-0.526) and 0.437 (0.355-0.524), respectively ($p < .001$). The Wilcoxon test also detected significant differences in reproducibility against B: features computed with a fixed R of 3mm were more reproducible than those computed with a fixed R of 1mm (Figure 3B), with median LCL (interquartile range) of 0.929 (0.853-0.988) and 0.833 (0.706-0.946), respectively, (p <.001).

Analogous to repeatability, reproducibility was unaffected by primary tumor (Figure E5 [online]), while differences were observed between liver and lung lesions (Table 2). This was true both in terms of number (Table E8 [online]) and type (Figures 3C, 3D) of reproducible RF. RF computed from lung lesions were more reproducible against both computation parameters than from liver lesions (p<.001), especially for features belonging to GLCM and GLRLM classes. Figure E6 (online) displays repeatability results for all features in the four reproducibility experiments. All statistical details related to reproducibility are reported in Table E9 (online).

Identification of Precise Features for Liver and Lung lesions

The LCL ≥0.50 threshold was chosen to remove non-precise features without over-eliminating potentially informative features with moderate precision. The identification was carried out for lung and liver lesions separately based on precision results. The identification yielded a total of 25 precise RF for liver and lung lesions, separately (Tables E10 and E11 [online]). We added the Neighborhood Gray-Tone-Difference-Matrix (NGTDM) Coarseness feature for both lesions for being among the top 3 most repeatable and reproducible against B (extended explanation in Appendix E7), resulting in 26 precise RF (Table 3. Figure 4 displays the results obtained in the three experiments for all features.

Imaging Habitats Computed with Precise Features Show Increased Stability

Figure 5B shows an example of the resulting habitats for one liver lesion. For that lesion, the heatmaps displaying significant correlations of precise RF and non-precise RF are available in Figure E7 and Figure E8 (online), respectively. The final lists of nonredundant RF and nonredundant-precise RF are reported in Table E12 (online).

The Wilcoxon signed-rank test revealed a statistically significant (p<.001) increase in habitat stability when habitats were computed with precise radiomics only (Figure 5B). This was true for habitats computed in both liver and lung lesions, observing a small effect size on both, with Cohen's *d* of 0.34 and 0.43, respectively. The median (interquartile range) DSC for habitat stability of habitats computed with non-precise RF was 0.532 (0.424-0.637) for lung and 0.587 (0.465-0.703) for liver lesions. For habitats computed with precise radiomics, the median scores were 0.601 (0.494-0.712) and 0.651 (0.52-0.784), respectively.

DISCUSSION

Computing robust and biologically meaningful tumor habitats (i.e., phenotypically similar regions within tumors) from clinical CT imaging could greatly advance non-invasive, 3D evaluation of tumor heterogeneity, one of the hallmarks of cancer resistance. However, to achieve an effective clinical translation of CT habitats, excellent robustness of the underlying imaging features is essential. In this study, we examined 3D radiomics' repeatability in a simulated test-retest scenario and reproducibility against kernel radius (R) and bin size (B), two significant computation parameters. Our findings demonstrated that 3D radiomics exhibit poor repeatability and reproducibility against R but excellent reproducibility against B (median LCL of the ICC of 0.442, 0.44, and 0.929, respectively). We identified different precise RF for lung and liver lesions, with primary tumor having no impact on precision. CT tumor habitats computed with precise feature subsets were more stable than those using all computed features (13% and 11% increase in median DSC in lung and liver lesions, P<.001). In an independent cohort, CT tumor habitats correlated both quantitatively and qualitatively with heterogeneity observed in mpMRI habitats and histology.

To the best of our knowledge, this is the first study evaluating the repeatability and reproducibility of CT 3D radiomics against R and B in the most common tumor locations, lung and liver. A direct comparison with other precision studies is therefore limited. Previous studies reported higher proportions of repeatable features (26-29), which could be attributed to differences in the number of RF, perturbation methods (27,28), and the use of phantoms instead of real patients (29). Our analysis found texture radiomics to be more precise than first-order (histogram) features, contrasting with earlier publications (11). This discrepancy might be due to histogram features being more influenced by outliers as they rely on absolute gray-level values. The high variability of 3D RF against R highlights the importance of caution in interpreting radiomics studies lacking detailed computation setting information. By providing this insight, we aim to minimize randomness in the radiomics workflow of future studies. Notably, we examined precision in the two common metastatic locations, lung and liver, and found differences between them, irrespective of primary tumor. One possible explanation for this difference could be the inherent differences in contrast-to-noise ratio between lung and liver lesions in CT imaging, which is generally higher in lung lesions. This indicates that performance of general radiomics models heavily depends on data characteristics, limiting generalizability to other tumor locations (30, 31). This finding provides a valuable foundation for future studies involving tumor habitats in heterogeneous cohorts and large-scale multicentric studies assessing cancer lesion heterogeneity.

We have demonstrated that the use of precise radiomics results in a more stable computation of CT habitats. Without a ground truth available, habitat stability was assessed on a voxel-by-voxel basis using the DSC between original-perturbed habitats. This local comparison might underestimate the global similarity between the compared habitats, potentially explaining the seemingly modest DSC values. Despite the inherent limitations of voxel-by-voxel comparisons, the large scale and design of our study provides a statistically significant and reliable answer to an ignored question: whether the preselection of RF by repeatability and reproducibility leads to an enhanced computation of imaging habitats in lung and liver lesions.

Our case study attempted to explore the biological relevance of CT habitats, inspired by previous studies that highlighted the value of quantitative MRI-derived habitats for characterizing tumor heterogeneity (15,16). We computed habitats independently in CT and mpMRI, observing that these imaging modalities consistently detected tumor heterogeneity (i.e. 2 or 3 habitats) in most lesions, reflecting similar underlying pathologic tissue compartments. Though conclusions are difficult to draw in view of our limited sample size, CT and mpMRI habitats may be capturing biologically relevant imaging phenotypes, potentially serving as non-invasive markers of cancer aggressiveness. This underscores the potential clinical utility of our approach, still in an exploratory context.

The generalizability of our results is subject to several key limitations. First, we focused on original features, without considering convolutional image filters like wavelet and Laplace of Gaussian filters. Convolutional filter-computed features have been shown to improve the predictive performance of radiomic signatures (32), yet their utility for habitat computation remains unknown and their standardization is still being developed (33). Second, we did not evaluate the impact of semi-automatic segmentation on 3D feature precision, but we warrant its assessment in future work with multiple delineation experiments. In addition, all segmentations were performed by one radiologist, which might introduce bias since features depend on contours. Although involving multiple delineators would provide a more comprehensive view of feature robustness, this was beyond the scope of our current study. Similarly, our study did not assess reproducibility across different scanners. This aspect, while crucial, is an area more explored in precision studies; thus, it was beyond the scope of the current study. Finally, our precision analysis was limited to CT data. Future studies are needed to study in detail the stability and robustness of other imaging modalities such as MRI or PET, as well as to investigate to what extent imaging phenotypes derived from different modalities overlap.

In summary, our comprehensive repeatability and reproducibility analysis identified two subsets of precise RF for effectively computing stable CT tumor habitats in lung and liver lesions. By employing these precise RF and using unsupervised clustering models, we demonstrated the ability to identify distinct tumor phenotypes in an exploratory analysis. CT tumor habitats correlated with biologically meaningful tumor aspects such as cellularity, vascularization, and necrosis, but further studies with larger sample sizes are needed to validate these findings. This approach to computing CT habitats holds great potential for studying intra-tumoral heterogeneity and monitoring cancer evolution throughout the course of the disease.

REFERENCES

1. Fidler IJ. Tumor heterogeneity and the biology of cancer invasion and metastasis. Cancer Res. 1978;38(9):2651–2660.

2. Swanton C. Intratumor heterogeneity: evolution through space and time. Cancer Res. 2012;72(19):4875–4882

3. Marusyk A, Janiszewska M, Polyak K. Intratumor Heterogeneity: The Rosetta Stone of Therapy Resistance. Cancer Cell. 2020;37(4):471–484.

4. McGranahan N, Swanton C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. Cancer Cell. 2015;27(1):15–26.

5. Xu H, Lv W, Feng H, et al. Subregional Radiomics Analysis of PET/CT Imaging with Intratumor Partitioning: Application to Prognosis for Nasopharyngeal Carcinoma. Mol. Imaging Biol. 2020;22(5):1414–1426.

6. Vargas HA, Veeraraghavan H, Micco M, et al. A novel representation of inter-site tumor heterogeneity from pre-treatment computed tomography textures classifies ovarian cancers by clinical outcome. Eur. Radiol. 2017;27(9):3991–4001.

7. Napel S, Mu W, Jardim-Perassi BV, et al. Quantitative imaging of cancer in the postgenomic era: Radio(geno)mics, deep learning, and habitats. Cancer. 2018;124(24):4633–4649.

8. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat. Rev. Clin. Oncol. 2017;14(12):749–762.

9. Kessler LG, Barnhart HX, Buckler AJ, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat. Methods Med. Res. 2015;24(1):9–26.

10. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology Standards for Quantitative Imaging Biomarkers. Radiology. 2015;277(3):813–825. Available at: http://dx.doi.org/10.1148/radiol.2015142202.

11. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. Int J Radiat Oncol Biol Phys. 2018;102(4):1143–1158.

12. Pfaehler E, Zhovannik I, Wei L, et al. A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. Physics and Imaging in Radiation Oncology. 2021. p. 69–75. doi:

10.1016/j.phro.2021.10.007.

13. Ng F, Kozarski R, Ganeshan B, et al. Assessment of tumor heterogeneity by CT texture analysis: can the largest cross-sectional area be used as an alternative to whole tumor analysis? Eur. J. Radiol. 2013;82(2):342–348.

14. Xu L, Yang P, Yen EA, et al. A multi-organ cancer study of the classification performance using 2D and 3D image features in radiomics analysis. Phys. Med. Biol. 2019;64(21):215009.

15. Jardim-Perassi BV, Huang S, Dominguez-Viqueira W, et al. Multiparametric MRI and Coregistered Histology Identify Tumor Habitats in Breast Cancer Mouse Models. Cancer Research. 2019. p. 3952–3964. doi: 10.1158/0008-5472.can-19-0213.

16. Divine MR, Katiyar P, Kohlhofer U, Quintanilla-Martinez L, Pichler BJ, Disselhorst JA. A Population-Based Gaussian Mixture Model Incorporating 18F-FDG PET and Diffusion-Weighted MRI Quantifies Tumor Tissue Classes. J Nucl Med. 2016;57(3):473–479.

17. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228–247.

18. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn. Reson. Imaging. 2012;30(9):1323–1341.

19. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. Sci. Rep. 2019;9(1):614.

20. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic

Phenotype. Cancer Res. 2017;77(21):e104–e107.

21. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology. 2020;295(2):328–338.

22. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J. Chiropr. Med. 2016;15(2):155–163.

23. Hinkle DE, Wiersma W, Jurs SG. Applied Statistics for the Behavioral Sciences. Houghton Mifflin; 2003.

24. Schwarz G. Estimating the Dimension of a Model. Ann. Stat. 1978;6(2):461–464.

25. Cohen J. A power primer. Psychol Bull. 1992;112(1):155–159.

26. Bernatowicz K, Grussu F, Ligero M, et al. Robust imaging habitat computation using voxel-wise radiomics features. Sci. Rep. 2021;11(1):20133.

27. Jha AK, Mithun S, Jaiswar V, et al. Repeatability and reproducibility study of radiomic features on a phantom and human cohort. Sci. Rep. 2021;11(1):2055.

28. Mottola M, Ursprung S, Rundo L, et al. Reproducibility of CT-based radiomic features against image resampling and perturbations for tumour and healthy kidney in renal cancer patients. Sci Rep. 2021;11(1):11542.

29. Berenguer R, Pastor-Juan MDR, Canales-Vázquez J, et al. Radiomics of CT Features May Be Nonreproducible

and Redundant: Influence of CT Acquisition Parameters. Radiology. 2018;288(2):407–415.

30. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-“how-to” guide and critical reflection. Insights Imaging. 2020;11(1):91.

31. Shur JD, Doran SJ, Kumar S, et al. Radiomics in Oncology: A Practical Guide. Radiographics. 2021;41(6):1717–1732.

32. Demircioğlu A. The effect of preprocessing filters on predictive performance in radiomics. Eur Radiol Exp. 2022;6(1):40.

33. IBSI. IBSI 2. Available at: https://theibsi.github.io/ibsi2/. Accessed January 16, 2023

TABLES

Table 1. Characteristics of Study Cohorts

	Colorectal	**Lung**	**Neuroendocrine**	**Mixed**
No. of Patients	75	85	86	85
No. of Images	215	124	86	180
No. of Lesions	1081	230	447	678
Age* (y)	64 (32-86)	65 (36-95)	61 (33-86)	68 (18-92)**
Sex, n (%)				
Male	33/75 (44%)	59/85 (69%)	42/86 (49%)	51/85 (60%)
Female	42/75 (56%)	26/85 (31%)	44/86 (51%)	34/85 (40%)

Note.—(*) Median (range)
(**) Missing information regarding 16 patients

Table 2. Median (IQR) Lower Confidence Limit of the Intraclass Correlation Coefficient for All Radiomics Features in Reproducibility Experiments

	Reproducibility against R				Reproducibility against B			
	Fixed B = 12HU		Fixed B = 25HU		Fixed R = 1mm		Fixed = 3mm	
Lesion type	Liver	Lung	Liver	Lung	Liver	Lung	Liver	Lung
Median Lower Confidence Limit (IQR)	0.422 (0.346-0.513)	0.573 (0.403-0.701)	0.407 (0.291-0.536)	0.573 (0.443-0.696)	0.805 (0.672-0.919)	0.929 (0.823-0.997)	0.921 (0.821-0.982)	0.967 (0.93-0.999)

Note.—B = bin size, HU = Hounsfield units, R = kernel radius

Table 3. Precise 3D Radiomics Features in Liver and Lung Lesions

Liver lesions	Lung lesions
FO_10Percentile	FO_90Percentile
FO_90Percentile	GLCM_Id
FO_Energy	GLCM_Idm
FO_Mean	GLCM_Imc1
FO_Median	GLCM_InverseVariance
FO_Minimum	GLCM_JointEntropy
FO_RootMeanSquared	GLDM_DependenceEntropy
GLCM_Autocorrelation	GLDM_DependenceNonUniformityNormalized
GLCM_JointAverage	GLDM_GrayLevelNonUniformity
GLCM_SumAverage	GLDM_LargeDependenceEmphasis
GLDM_DependenceEntropy	GLDM_LargeDependenceHighGrayLevelEmphasis
GLDM_GrayLevelNonUniformity	GLDM_SmallDependenceEmphasis
GLDM_HighGrayLevelEmphasis	GLDM_SmallDependenceHighGrayLevelEmphasis
GLDM_LargeDependenceLowGrayLevelEmphasis	GLRLM_GrayLevelNonUniformity
GLDM_LowGrayLevelEmphasis	GLRLM_LongRunEmphasis
GLDM_SmallDependenceHighGrayLevelEmphasis	GLRLM_LongRunHighGrayLevelEmphasis
GLRLM_GrayLevelNonUniformity	GLRLM_RunLengthNonUniformity
GLRLM_HighGrayLevelRunEmphasis	GLRLM_RunLengthNonUniformityNormalized
GLRLM_LongRunHighGrayLevelEmphasis	GLRLM_RunPercentage
GLRLM_LongRunLowGrayLevelEmphasis	GLRLM_RunVariance
GLRLM_LowGrayLevelRunEmphasis	GLRLM_ShortRunEmphasis
GLRLM_RunLengthNonUniformity	GLSZM_LargeAreaEmphasis
GLRLM_RunPercentage	GLSZM_LargeAreaHighGrayLevelEmphasis
GLRLM_RunVariance	GLSZM_ZonePercentage
GLRLM_ShortRunHighGrayLevelEmphasis	GLSZM_ZoneVariance
NGTDM_Coarseness	NGTDM_Coarseness

Note.— A precise radiomic feature was defined as lower confidence limit ≥ 0.50 in the repeatability and reproducibility experiments. FO = first-order; GLCM = Grey Level Co-occurrence Matrix features; GLDM = Grey Level Dependence Matrix; GLRLM = Grey Level Run Length Matrix; GLSZM = Grey Level Size Zone Matrix; NGTDM = Neighboring Grey Tone Difference Matrix Features.

Supplemental Material

Appendix E1. Independent Cohort for Biological Relevance

Image acquisition

13 patients with cancer in the liver (either primary or metastatic) from a prospective clinical trial were included. All patients underwent contrast-enhanced CT imaging on a General Electric (GE) scanner (STANDARD convolution kernel, KVP: 120) as well as abdominal MRI on two different MRI machines. These were a 1.5T Siemens Avanto and a 3T GE SIGNA Pioneer scanner. MRI data was acquired for each patient using only one of the two scanners, across multiple time points (only baseline scans were considered for this study). Salient details of the MRI protocol are reported below.

1.5T Siemens Avanto system

The protocol included high-resolution anatomical T2w and T1w scans, diffusion MRI and different spoiled gradient echo (SGrE) sequences, such as those for T1 mapping and dynamic contrast enhanced (DCE) MRI.

- Anatomical T2w scan: turbo spin echo, TE = 82 ms, TR = 4500 ms, turbo factor of 29, echo spacing 8.2 ms, NEX = 8, 2 concatenations, resolution of 1.4mm × 1.4mm, slice thickness of 5 mm, GRAPPA = 2, acquisition in free breathing.
- Anatomical T1w scan: turbo spin echo, TE = 6.3 ms, TR = 470 ms, turbo factor of 11, echo spacing 6.26 ms, NEX = 6, 6 concatenations, resolution of 1.4mm × 1.4mm, slice thickness of 5 mm, GRAPPA = 2, acquisition in free breathing.
- Diffusion MRI: single-shot twice-refocused spin echo EPI, b = {0, 50, 100, 400, 900, 1200, 1600} s/mm^2, TR = 7900 ms, averaging of 3 mutually-orthogonal directions, NEX = 2, 1 concatenation, resolution of 1.9mm × 1.9mm, slice thickness of 6 mm, SPAIR fat suppression, GRAPPA = 2, EPI factor 150, echo spacing 0.82 ms, each b-value acquired at TE = {93 ms, 105 ms, 120 ms}, acquisition in free breathing. Additionally, one b = 0 image at TE = 93 ms was acquired with reversed phase encoding polarity.
- SGrE for T1 mapping: FLASH, TE = 1.76 ms, TR = 4.59 ms, NEX = 1, 1 concatenation, resolution of 2.7mm

× 2.7mm, slice thickness of 6 mm, flip angles of {5°, 15°, 20°}, GRAPPA = 2, acquisition in free breathing.

- SGrE for DCE: same acquisition as for T1 mapping with fixed flip angle of 15°; 26 dynamic acquisitions with temporal resolution of 10s, Gadovist with dose of 0.5ml/Kg injected at 3ml/s followed by a bolus of physiological solution of 20ml at 3ml/s, injection delay of 10s, acquisition in free breathing.

3T GE SIGNA Pioneer system

The protocol included high-resolution anatomical T2w and T1w scans, diffusion MRI and different spoiled gradient echo (SGrE) sequences, as those for T1 mapping and DCE imaging.

- Anatomical T2w scan: fast spin echo, TE = 50 ms, TR = 3750 ms, turbo factor of 16, NEX = 2, resolution of 1.4mm × 1.4mm, slice thickness of 6 mm, respiratory-gated acquisition.
- Anatomical T1w scan: navigated SGrE LAVA-Flex providing water/fat images, TE = 2.60 ms, TR = 5.38 ms, resolution of 1.4mm × 1.4mm, slice thickness of 6 mm, flip angle of 12°, acquisition in free-breathing after liver motion measurement.
- Diffusion MRI: single-shot pulsed gradient spin echo EPI, b = {0, 50, 100, 400, 900, 1200, 1500} s/mm^2, TR = 3500ms, averaging of 3 mutually-orthogonal directions, NEX = 2, resolution of 2.4mm × 2.4mm, slice thickness of 6 mm, ASPIR fat suppression, ASSET = 2, echo spacing 0.80 ms, each b-value acquired at TE = {75 ms, 90 ms, 105 ms}, respiratory-gated acquisition.
- SGrE for T1 mapping: LAVA, TE = 1.2 ms, TR = 2.72 ms, NEX = 1, resolution of 2.4mm × 2.4mm; slice thickness of 6 mm; flip angles of {5°, 10°, 15°}, ASSET = 2, acquisition of two separate images in breath-hold, acquisition of the vendor’s B1 map.
- SGrE for DCE: same acquisition as for T1 mapping with fixed flip angle of 12°; 26 dynamic acquisitions with temporal resolution of 10s, Clariscan 0.5 mmol/ml with dose of 0.2ml/Kg injected at of 0.5ml/kg at 3ml/s followed by a bolus of physiological solution of 20ml at 3ml/s, injection delay of 10s, acquisition in free breathing.

MRI pre-processing

- Diffusion MRI: MP-PCA denoising, Gibbs unringing, motion correction based on affine co-registration, EPI distortion correction when a reversed phase encoding scan was available.
- SGrE: MP-PCA denoising, motion correction based on affine co-registration.
- Anatomical imaging: N4 bias field correction based on ANTs and affine co-registration of the T1w scan to the T2w scan.
- Lesion segmentations: lesions were outlined manually by an experienced radiologist (RPL) on the T2w anatomical scan, aided by visual inspection of all MRI contrasts.
- Co-registration: a non-linear transformation warping diffusion and SGrE image spaces to the T2w anatomical scan was calculated with ANTs.

MRI parametric map calculation

The following biophysical models were fitted to the MRI data.

- Diffusion MRI: a two-pool diffusion-T2 relaxation model capturing signal contributions from vascular water (characterized by intra-voxel incoherent motion (IVIM)) and non-vascular (tissue) water. The following metrics were obtained: apparent diffusion coefficient of tissue and vascular water (ADCt, ADCv); T2 of tissue and vascular water ($T2t$, $T2v$), vascular fraction (fv), and tissue kurtosis excess (Kt). $T2v$ was excluded from downstream processing due to lack of sensitivity of the acquisition protocol to long T2 components. The model was fitted with custom-written python code.
- T1 mapping: mono-exponential quantitative T1 fitting on variable flip angle imaging, correcting the flip angle of the SGrE signal expression with the B1 map when available. The model was fitted using the freely available MyRelax python toolbox (https://github.com/fragrussu/MyRelax/blob/master/myrelax/getT1VFA.py).
- DCE: variational Bayesian fitting of a two-compartment Toft's exchange model with FSL FABBER (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FABBER). An Orton's arterial input function was used and T1 was fixed voxel-by-voxel to the value estimated in previous variable flip angle T1 mapping. DCE fitting provided voxel-wise estimates of the volume transfer constant K_{trans} (capillary permeability constant), plasma and extra-

cellular volume fractions (v_p and v_e)

After model fitting, all parametric maps were co-registered to the T2-weighted anatomical image using the non-linear warp previously estimated with ANTs, enabling downstream voxel-wise habitat computation with the segmented lesions.

Digital Pathology

An ultra-sound guided biopsy was taken by an experienced radiologist from one of the lesions imaged with MRI. The histological material was processed as per standard procedures (e.g., formalin fixation, paraffin-embedding, HE staining) and a digital image of the HE-stained image was acquired using a Hamamatsu C9600-12 scanner (resolution: 0.45 µm). An experienced pathologist (SS) inspected the images and outlined areas containing cancer cells, as well as areas of fibrosis, necrosis, and any other tissues from outside the biopsied lesion.

Appendix E2. Image perturbation

Image perturbation was carried out in three ways: rotation, translation, and noise addition. While the first two emulate changes in patient positioning, the latter represents the noise present in different voxel intensities in CT images. Perturbations were performed as described in (19, 24), where the authors demonstrated that the combination of these three perturbations simulate the retest scenario. Briefly, we added Gaussian noise (mean 0, standard deviation as present in the image); for translation, we shifted the voxel grid by a fraction of the image voxel spacing following; finally, we rotated the image around the z-axis by an angle of 0.5°.

Appendix E3. Feature Computation

To compute voxel-wise (3D) features, the software first picks one voxel of interest, defines a cubic kernel in the neighborhood of the voxel of radius R and size $\mathbf{2} \times \boldsymbol{R} + \mathbf{1}$ (thus, R=1mm creates a kernel size 3x3x3 and R=3mm creates one of 7x7x7 dimensions), calculates the features in the kernel (only considering voxels lying inside the cube), and reports the calculated value as the new voxel of a new image. Thus, each voxel-wise feature is a SimpleITK image (i.e. library used by PyRadiomics [34, 35]). Figure E1 provides two examples of computed features. Table E4 describes the full list of computed features. All relevant image processing and feature computation parameters are reported in Table E5.

Appendix E4. Habitat Computation

To take into account intravoxel heterogeneity, we decided to choose a probabilistic model, Gaussian Mixture Models (GMMs), for clustering rather than a deterministic approach. GMMs, which have been previously used in similar contexts (36, 37), are generative probabilistic models that find a mixture of multiple Gaussian probability distributions that best fit the data. The Expectation-Maximization (EM) algorithm is used to estimate the model parameters (38). A GMM is represented by the following formula:

$$P\ (x) = \sum (\pi_k\ N\ (x \mid \mu_k, \Sigma_k)$$

where

- **P(x) :** probability density of the data point x
- $\boldsymbol{\pi_k}$**:** mixing coefficient for the kth Gaussian component
- $\boldsymbol{N\ (x \mid \mu_k, \Sigma_k)}$**:** kth Gaussian component with mean $\boldsymbol{\mu_k}$ and covariance matrix $\boldsymbol{\Sigma_k}$

To determine the optimal number of habitats (k), we used the Bayesian Information Criterion (BIC). The formula for BIC is:

$$BIC = -2\log(L) + d\ log(n)$$

where

- **L :** likelihood of the data given the model
- **d:** number of parameters

- **n:** number of data points

The BIC score is a measure of the trade-off between model complexity and goodness of fit. It penalizes models with more parameters, such as GMMs with more clusters. In general, lower BIC scores indicate better model fit. However, depending on data characteristics, a clear minimum in BIC scores might not be observed and thus, the gradient can be used to determine the optimal number of clusters. This was our case (for an example see Figure E2 [online]). Thus, we performed a GMM fit for different values of clusters (k): {2, 3, 4 and 5}. The maximum number of 5 clusters was determined by being the maximum number of tissue types observed in histology by an experienced pathologist. The optimal value of k was the one where the change in BIC score with respect to k was maximal, which was an indication that adding more clusters after that point does not improve the model fit significantly. A cluster number was automatically selected by BIC using the precise original radiomics data and was given as a parameter to the GMM model to compute imaging habitats in both the original and perturbed data. GMM was implemented using Python package scikit-learn (v1.0.2) with a random seed of 123, and default parameters (except for the number of clusters), specifically maximum iteration of 100, convergence threshold of 10^{-3}, full covariance type and initialization with kmeans.

In addition, The Hungarian algorithm (also known as the Kuhn-Munkres algorithm) (39), was used to match habitats between original and perturbed data. The Hungarian algorithm is a combinatorial optimization algorithm that solves the assignment problem in polynomial time. It finds an optimal one-to-one matching between two sets by minimizing the total cost (in our case, the difference in cluster assignments).

Finally, to quantify habitat stability, we computed the Dice Similarity Coefficient (DSC) (40) between original and perturbed habitats for each habitat within a lesion, across all lesions. The DSC is a widely used metric for evaluating the overlap between two sets, with a higher DSC indicating greater similarity.

All codes are publicly available at https://github.com/radiomicsgroup/precise-habitats.

Appendix E5. Case Study: imaging habitats as potential indicators of intratumoral heterogeneity

Methods

We evaluated the biological relevance of CT imaging habitats in an independent cohort of 13 patients with CT, mpMRI (including anatomical, diffusion-weighted and perfusion MRI), and digitized Hematoxylin-Eosin (HE) images from biopsy. Habitats were computed using GMM and BIC in the same fashion as described above on both CT and mpMRI. Specifically, for CT, habitats were computed with the subset of precise liver features, eliminating redundant ones as earlier described. For mpMRI, habitats were computed by voxel-wise clustering of nine mpMRI maps, warped non-linearly on top of a high-resolution T2-weighted anatomical scan. The mpMRI metrics were tissue T2 (T2t), longitudinal relaxation time (T1), tissue/vascular apparent diffusion coefficient (ADCt, ADCv), tissue apparent kurtosis (Kt) and vascular fraction (fv) from the diffusion-weighted MRI, capillary permeability constant (Ktrans), extra-cellular extra-vascular volume fraction (ve), and plasma volume fraction (vp) from perfusion MRI. More information regarding the maps is available on Table E6 (online). The number of CT habitats and mpMRI habitats, as selected by BIC, were compared quantitatively for each lesion. The distributions of both CT and mpMRI habitats were also compared by visual alignment. Lastly, we conducted an imaging-histology qualitative evaluation, without spatial alignment, to assess whether the variety of imaging habitats could be potentially correlated with the diversity of tissue phenotypes identified by a pathologist during histologic examination.

Results

The GMM model with BIC yielded the same number of habitats for 10 out of 13 lesions (Figure E9 [online]). Our qualitative evaluation, focusing on the visual inspection of mpMRI and CT habitats and their potential relation to the diversity of tissue phenotypes identified in histology, showed comparable distributions. This is illustrated in Figure E10 for one representative patient (liver metastasis of melanoma). The habitat depicted in blue, captured in both CT and mpMRI and visually compatible with an area of necrosis, shows higher T2t, T1 and ADCt (compatible with lower cellularity), as well as lower fv, Ktrans, ve and vp (compatible with lower vascularization), as compared with the rest of the lesion. The digital pathology - which was used to qualitatively observe the diversity of tissue types rather than for precise spatial correlations - confirmed the presence of necrosis (i.e., non-viable cells and lack of vascularization) within the core of the lesion. Another example is shown in Figure E11 (online).

Appendix E6. Statistical Analysis

An Intraclass Correlation Coefficient (ICC) value of 1 indicates that a feature is highly repeatable/reproducible whereas a value of 0 implies no reliability. Negative ICC values were truncated at 0 as proposed and done previously (41, 42). The ICC is calculated by mean squares obtained through the analysis of variance (ANOVA). In this study, we use two versions of the ICC that are based on a two-way mixed effects ANOVA model, following Koo's guidelines. Below we describe the formulas used to compute the ICC formulas. More information regarding such formulas can be found in the highly cited paper from McGraw and Wong.

To compute the ANOVA model let's consider a dataframe with dimensions $\boldsymbol{n} \times \boldsymbol{k}$ dataframe where $\boldsymbol{n}$ is the total number of voxels (rows) for one region of interest (ROI) and $\boldsymbol{k}$ is the total number of conditions or measurements (columns). In our case, $\boldsymbol{k}$=2. For repeatability the two conditions are original-perturbed (test-retest) and for reproducibility against kernel size the two conditions are computation with radius kernel 1mm or radius kernel 3mm (or bin size 12HU or 25HU in the case of reproducibility against bin size). Each voxel measurement is indexed as $\boldsymbol{Y_{ij}}$ where *i* denotes the voxel (*i* = 1, … *n)* and *j* denotes the measurement under the repeatability/reproducibility condition (*j = 1 ... k)*. We define the following concepts:

- $\boldsymbol{\bar{Y}_i}$**:** mean of all voxel values in a column

$$\boldsymbol{\bar{Y}_i = \frac{\sum_{j=1}^{k} Y_{ij}}{k}}$$

- $\boldsymbol{\bar{Y}_j}$: mean of all voxel values in a column

$$\overline{Y}_j = \frac{\sum_{i=1}^{n} Y_{ij}}{n}$$

- μ: mean of all values (also called *grand mean*)

$$\mu = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n} Y_{ij}}{n * k}$$

- ${\sigma_w}^2$: Within-voxel variance, the estimated variance of repeated measurements

$${\sigma_w}^2 = \frac{\sum_{j=1}^{k} (Y_{ij} - \overline{Y}_i)^{\,2}}{k - 1}$$

- σ_w : Within-voxel standard deviation, the standard deviation we get if we measure the voxel multiple times. Calculated by averaging the within-subject sample variances. Since we have a variance per voxel and we can't meaningfully take the average of a list of standard deviations, we first calculate the variance for each voxel, and then compute the average of those, and finally square root that mean variance (43, 44).

$$\sigma_w = \sqrt{\frac{\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} (Y_{ij} - \overline{Y}_i)^{\,2}}{k - 1}}{n}}$$

The degrees of freedom, sum squares and mean square expectations that correspond to a two-way mixed ANOVA model are summarized below.

Source of Variation	Degrees of freedom	Sum Squares	Mean Square Expectations
Conditions (columns)	dfc = k -1	$SSC = \sum_{j=1}^{k} n \times (\overline{Y}_j - \mu)^{\,2}$	$MSC = \frac{SSC}{dfc \times n}$

Voxels (rows)	dfr = n -1	$SSR = \sum_{i=1}^{n} k \times (\bar{Y}_i - \mu)^{2}$	$MSR = \frac{SSR}{dfr}$
Total	-	$SST = \sum_{j=1}^{k} \sum_{i=1}^{n} (Y_{ij} - \mu)^{2}$	-
Error (or residual)	dfe = (n -1)(k -1)	$SSE = SST - SSC - SSR$	$MSE = \frac{SSE}{dfe}$

MSC: mean square columns, MSR: mean square rows, MSE: mean square error, SSC= sum of squares columns, SSR=sum of squares rows, SST= sum of squares total, SSE= sum of squares error, dfc= degrees of freedom columns, dfr=degrees of freedom rows, dfe=degrees of freedom errors

We compute the two versions of ICC for repeatability and reproducibility:

- **Repeatibility ICC(3A,1):** ICC based on single-measurement, absolute-agreement, two-way mixed-effects model.

$$ICC\ (3A, 1) = \frac{MSR - MSE}{MSR + dfc\ \times\ MSE + \frac{k}{n} \times (MSC - MSE)}$$

- **Reproducibility ICC(3C,1):** ICC based on single-measurement, consistency, two-way mixed-effects model.

$$ICC(3C, 1) = \frac{MSR - MSE}{MSR + dfc \times\ MSE}$$

We compute the lower bound of the 95% CI of the ICC (LCL) and the upper bound (UCL):

$$\boldsymbol{LCL} = \frac{\frac{\boldsymbol{FR}}{\boldsymbol{F}} - \boldsymbol{1}}{\frac{\boldsymbol{FR}}{\boldsymbol{F}} + \boldsymbol{k} - \ \boldsymbol{1}} \qquad \boldsymbol{UCL} = \frac{(\boldsymbol{FR} \ \times \ \boldsymbol{F}) \ - \boldsymbol{1}}{(\boldsymbol{FR} \ \times \ \boldsymbol{F}) + \boldsymbol{k} - \ \boldsymbol{1}}$$

Where F is the (1-$\frac{\alpha}{2}$) x 100th percentile of the F distribution with n-1 numerator degrees of freedom and (n-1)(k-1) denominator degrees of freedom and FR is the F-statistic for voxels computed as: $\boldsymbol{FR} = \frac{\boldsymbol{MSR}}{\boldsymbol{MSE}}$.

Custom codes used to calculate ICC (3A,1) and ICC (3C,1)**,** based on Nipype's (45) module ICC (v1.8.5) and approved by an statistician (VN) are available at https://github.com/radiomicsgroup/precise-habitats.

Appendix E7. NTGDM Coarseness is a Precise Feature

NGTDM (Neighborhood Gray-Tone-Difference Matrix) coarseness describes the roughness (i.e. how fine or coarse) the texture of an image is. In the radiomics literature, evidence has been found regarding its usefulness to characterize heterogeneity and predict progression-free survival in oncology (46).

In our study, we identified precise features by linking repeatability and reproducibility results. That is, for every feature, we considered results obtained in the three relevant experiments: repeatability (setting R3B12), reproducibility against R (fixed B=12HU), and reproducibility against B (fixed R=3mm). A feature was selected as precise if it presented LCL ≥ 0.50 (i.e. moderate, good or excellent repeatability/reproducibility) in the three experiments. NGTDM Coarseness presented excellent repeatability and reproducibility against bin size, but was not selected as precise as it presented poor reproducibility against kernel radius. However, by the nature of its definition, the poor reproducibility against kernel radius is acceptable: the feature captures the distribution of differences in gray-tone values between pairs of neighboring pixels. Considering the feature's excellent results in two out of three experiments, its potential usefulness and in light of the fact that we were already being stringent, first by using LCL rather than ICC and second by linking results of three different experiments, , we decided to include it as a precise feature for both liver and lung lesions.

Supplemental References

34. Yaniv Z, Lowekamp BC, Johnson HJ, et al. Correction to: SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. J. Digit. Imaging. 2019;32(6):1118.

35. Yaniv Z, Lowekamp BC, Johnson HJ, et al. Correction to: SimpleITK Image-Analysis Notebooks: a Collaborative Environment for Education and Reproducible Research. J. Digit. Imaging. 2019;32(6):1118.

36. Jardim-Perassi BV, Huang S, Dominguez-Viqueira W, et al. Multiparametric MRI and Coregistered Histology Identify Tumor Habitats in Breast Cancer Mouse Models. Cancer Research. 2019. p. 3952–3964. doi: 10.1158/0008-5472.can-19-0213.

37. Chen J, Milot L, Cheung HMC, Martel AL. Unsupervised Clustering of Quantitative Imaging Phenotypes Using Autoencoder and Gaussian Mixture Model. Medical Image Computing and Computer Assisted Intervention – MICCAI 2019. Springer International Publishing; 2019. p. 575–582.

38. Bishop CM. Pattern Recognition and Machine Learning. Springer New York;

39. Kuhn HW. The Hungarian method for the assignment problem. Nav Res Logist Q. Wiley; 1955;2(1-2):83–97.

40. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol. 2004;11(2):178–189.

41. Bartko JJ. On various intraclass correlation reliability coefficients. Psychol. Bull. 1976;83(5):762–765.

42. Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. Eur. Radiol. 2020;30(11):6241–6250.

43. FMRI Biomarker Committee. Indices of Repeatability, Reproducibility, and Agreement [Internet]. Quantitative Imaging Biomarkers Alliance (QIBA); 2013.Available from: http://qibawiki.rsna.org/images/e/e3/FMRITechnicalPerformanceIndices042613.pdf

44. Ye S, Lim JY, Huang W. Statistical considerations for repeatability and reproducibility of quantitative imaging biomarkers. *Microbiologyopen*. 2022;4(1):20210083.

45. Esteban O, Markiewicz CJ, Burns C, et al. nipy/nipype: 1.8.3.; 2022. Available at: https://zenodo.org/record/6834519.

46. Gupta H, Deek MP, McNutt TR, Lee J, Quon H, Sheikh K. Predicting Progression-Free Survival Using Primary and Nodal Radiomic Features in Head and Neck Cancer. Int J Radiat Oncol Biol Phys. Elsevier; 2021;111(3):e130–e131.

Supplemental Tables

Table E1. Total number of patients, images, and lesions per cohort and lesion location.

Primary tumor	Patients	Images	Lesions

	Liver	**Lung**	**Liver**	**Lung**	**Liver**	**Lung**
Colorectal	63	12	186	29	959	122
Lung	13	72	22	102	89	141
Neuroendocrine	86	0	86	0	447	0
Mixed	44	41	93	87	366	312
Total	**206**	**125**	**387**	**218**	**1861**	**575**
	331		**605**		**2436**	

Table E2. List of primary tumor types included within the mixed cohort.

Primary tumor	**Patients**		**Images**		**Lesions**	
	Liver	**Lung**	**Liver**	**Lung**	**Liver**	**Lung**
Adrenal	3	0	5	0	35	5
Biliary Tract	11	5	24	11	71	66

Bladder	3	3	5	5	41	20
Bone	0	1	0	3	0	21
Breast	4	2	9	6	32	6
Cervix	2	2	3	3	21	21
Esophagus	1	2	2	4	8	9
Head&Neck	2	4	4	11	13	13
Kidney	1	2	2	3	4	18
Liver	2	1	3	1	11	1
Ovary	1	2	2	4	4	26
Pancreas	2	0	4	0	12	0
Penis	1	0	3	0	9	0
Skin	6	11	13	19	68	39
Stomach	4	0	10	0	29	0
Thymus	1	0	4	0	8	0
Thyroid	0	6	0	17	0	52
Total	**44**	**41**	**93**	**87**	**366**	**312**
	85		**180**		**678**	

Table E3. Image acquisition parameters per cohort.

	Colorectal (n=215)	**Lung (n=124)**	**Neuroendocrine (n=86)**	**Mixed (n=180)**
Manufacturers SIEMENS/PHILIPS/	138/58/9/10	63/40/0/21	22/35/6/23	144/23/3/10

TOSHIBA/GENERAL ELECTRIC				
Tube Voltage (kVP) 100/110/120/130/140/unknown	25/17/161/0/0/12	6/1/105/3/0/9	10/3/70/1/2/0	14/7/158/0/1/0
Reconstruction kernel SOFT/STANDARD/B	1/7/96	14/7/43	1/19/36	2/11/21
B20f/B30f/B31f	16/38/12	11/23/0	2/2/4	9/112/4
I31s/I50s/unkown	15/6/20	0/0/22	0/0/19	0/0/21
Slice thickness (mm)*	2.0 [2.00-5.00]	2.5 [2.0-5.0]	2.0 [2.0-3.0]	5.0 [1.00-5.00]
Pixel spacing (mm)*	0.92 [0.77-0.98]	0.91 [0.81-0.98]	0.75 [0.70-0.82]	0.98 [0.82-0.98]

(*) Median [IQR]

Table E4. List of Radiomics features analyzed in this study.

Class	Feature	Class	Feature

First order	1. 10Percentile 2. 90Percentile 3. Energy 4. Entropy 5. InterquartileRange 6. Kurtosis 7. Maximum 8. MeanAbsoluteDeviation 9. Mean 10. Median 11. Minimum 12. Range 13. RobustMeanAbsoluteDeviation 14. RootMeanSquared 15. Skewness 16. TotalEnergy 17. Uniformity 18. Variance
GLCM	1. Autocorrelation 2. ClusterProminence 3. ClusterShade 4. ClusterTendency 5. Contrast 6. Correlation 7. DifferenceAverage 8. DifferenceEntropy 9. DifferenceVariance 10. Id 11. Idm 12. Idmn 13. Idn 14. Imc1 15. Imc2 16. InverseVariance 17. JointAverage 18. JointEnergy 19. JointEntropy 20. MCC 21. MaximumProbability 22. SumAverage 23. SumEntropy 24. SumSquares
NGTDM	1. Busyness 2. Coarseness 3. Complexity 4. Contrast 5. Strength
GLRLM	1. GrayLevelNonUniformity 2. GrayLevelNonUniformityNormalized 3. GrayLevelVariance 4. HighGrayLevelRunEmphasis 5. LongRunEmphasis 6. LongRunHighGrayLevelEmphasis 7. LongRunLowGrayLevelEmphasis 8. LowGrayLevelRunEmphasis 9. RunEntropy 10. RunLengthNonUniformity 11. RunLengthNonUniformityNormalized 12. RunPercentage 13. RunVariance 14. ShortRunEmphasis 15. ShortRunHighGrayLevelEmphasis 16. ShortRunLowGrayLevelEmphasis
GLSZM	1. GrayLevelNonUniformity 2. GrayLevelNonUniformityNormalized 3. GrayLevelVariance 4. HighGrayLevelZoneEmphasis 5. LargeAreaEmphasis 6. LargeAreaHighGrayLevelEmphasis 7. LargeAreaLowGrayLevelEmphasis 8. LowGrayLevelZoneEmphasis 9. SizeZoneNonUniformity 10. SizeZoneNonUniformityNormalized 11. SmallAreaEmphasis 12. SmallAreaHighGrayLevelEmphasis 13. SmallAreaLowGrayLevelEmphasis 14. ZoneEntropy 15. ZonePercentage 16. ZoneVariance
GLDM	1. DependenceEntropy 2. DependenceNonUniformity 3. DependenceNonUniformityNormalized 4. DependenceVariance 5. GrayLevelNonUniformity 6. GrayLevelVariance 7. HighGrayLevelEmphasis 8. LargeDependenceEmphasis 9. LargeDependenceHighGrayLevelEmphasis 10. LargeDependenceLowGrayLevelEmphasis 11. LowGrayLevelEmphasis 12. SmallDependenceEmphasis 13. SmallDependenceHighGrayLevelEmphasis 14. SmallDependenceLowGrayLevelEmphasis

93 voxel-wise features were computed. However, 91 were analyzed after excluding GLCM_MCC (for having missing values in many cases due to memory error) and FirstOrder_TotalEnergy (for being equal to FirstOrder_Energy for constant kernel size during computation). Feature definitions are available in the IBSI reference manual (19).

Table E5. Image processing and radiomics feature computation parameters.

Image Processing	
Software	PyRadiomics v3.0.1, installed in Python 3.7.10
Bounding box	Defined by the segmentation, extended by default padding distance.
Resampled voxel spacing (mm)	1 x 1 x 1
Image interpolation method	B-spline
Intensity rounding	None
ROI interpolation method	Nearest neighbor
Resegmentation	None
Feature Computation	
Kernel radius	1 / 3mm
Discretization (fixed bin size)	12/25HU
Image filter	None
maskedKernel	True (only voxels in the kernel that were also segmented in the ROI were used for calculation)
Initvalue	NaN (voxels outside ROI were considered as transparent)
Distance weighting for GLCM, GLRLM, NGTDM	No weighting
GLCM Symmetry	Symmetric
GLCM distance, GLSZM linkage distance, GLDZM linkage distance, NGTDM distance	Chebyshev distance δ = 1
NGTDM coarseness	Coarseness parameter α = 0

Table E6. Mulitparametric (mpMRI) biomarkers used to compute mpMRI-habitats for biological relevance assessment.

Biomarker	Computed from	Units	Description
Tissue Apparent Diffusion Coefficient (ADCt)	Diffusion-relaxation MRI	$\mu m^2\ ms^{-1}$	Apparent diffusivity of water in the tissue compartment (i.e., excluding vasculature)
Vascular Apparent Diffusion Coefficient (ADCv)	Diffusion- relaxation MRI	$\mu m^2\ ms^{-1}$	Apparent diffusivity of water in the vascular compartment (i.e., from fast pseudo-diffusion sensitive to the intra-vascular incoherent motion (IVIM) effect)
Tissue Apparent Kurtosis Excess (Kt)	Diffusion- relaxation MRI	Dimensionless	Metric quantifying departures from non-Gaussian diffusion due to diffusion heterogeneity or diffusion restriction
Vascular signal fraction (fv)	Diffusion- relaxation MRI	Normalised	Amount of signal arising from the vascular compartment (i.e., from fast pseudo-diffusion sensitive to the intra-vascular incoherent motion (IVIM) effect
Tissue Transverse Relaxation time (T2t)	Diffusion- relaxation MRI	ms	Transverse relaxation time of the tissue compartment (i.e., excluding vasculature)
Total Longitudinal Relaxation Time (T1)	Variable flip angle spoiled gradient echo MRI	ms	Total longitudinal relaxation time
Capillary Permeability Constant (Ktrans)	Dynamic Contrast Enhanced MRI	min^{-1}	Influx mass transfer rate of contrast, measuring permeability/vascular leak .
Plasma volume fraction (vp)	Dynamic Contrast Enhanced MRI	Normalised	Volume of plasma per unit volume of tissue
Extra-cellular, extra-vascular voluem fraction (ve)	Dynamic Contrast enhanced MRI	Normalised	Volume of tissue extra-cellular, extra-vascular space

Table E7. Number of repeatable features per lesion location for all repeatability experiments

	R1B12	R1B25	R3B12	R3B25

	Liver	Lung	Liver	Lung	Liver	Lung	Liver	Lung
Poor	87/91 (95.6%)	82/91 (90.1%)	88/91 (96.7%)	85/91 (93.4%)	55/91 (60.4%)	60/91 (65.9%)	55/91 (60.4%)	66/91 (72.5%)
Moderate	4/91 (4.4%)	8/91 (8.8%)	3/91 (3.3%)	6/91 (6.6%)	30/91 (33.0%)	29/91 (31.9%)	29/91 (31.9%)	23/91 (25.3%)
Good	0/91 (0.0%)	1/91 (1.1%)	0/91 (0.0%)	0/91 (0.0%)	5/91 (5.5%)	1/91 (1.1%)	7/91 (7.7%)	2/91 (2.2%)
Excellent	0/91 (0.0%)	0/91 (0.0%)	0/91 (0.0%)	0/91 (0.0%)	1/91 (1.1%)	1/91 (1.1%)	0/91 (0.0%)	0/91 (0.0%)

Table E8. Number of reproducible features against R and B per lesion location for all reproducibility experiments

	Reproducibility against kernel radius				**Reproducibility against bin size**			

	Fixed B = 12HU		Fixed B = 25HU		Fixed R = 1mm		Fixed = 3mm	
	Liver	Lung	Liver	Lung	Liver	Lung	Liver	Lung
Poor	65/91 (71.4%)	34/91 (37.4%)	63/91 (69.2%)	35/91 (38.5%)	11/91 (12.1%)	4/91 (4.4%)	6/91 (6.6%)	0/91 (0.0%)
Moderate	17/91 (18.7%)	47/91 (51.6%)	20/91 (22.0%)	45/91 (49.5%)	29/91 (31.9%)	12/91 (13.2%)	11/91 (12.1%)	5/91 (5.5%)
Good	9/91 (9.9%)	10/91 (11.0%)	8/91 (8.8%)	11/91 (12.1%)	23/91 (25.3%)	23/91 (25.3%)	24/91 (26.4%)	8/91 (8.8%)
Excellent	0/91 (0.0%)	0/91 (0.0%)	0/91 (0.0%)	0/91 (0.0%)	28/91 (30.8%)	52/91 (57.1%)	50/91 (54.9%)	78/91 (85.7%)

Table E9. Wilcoxon signed rank test results for reproducibility

	Reproducibility against R			Reproducibility against B		
	Fixed B12 vs Fixed B25	Lung vs Liver (fixed B=12HU)	Lung vs Liver (fixed B=25HU)	Fixed R=1mm vs Fixed R=3mm	Lung vs Liver (fixed R=1mm)	Lung vs Liver (fixed R=3mm)
Z	777.5	856	602.5	72.5	84	167
p	<.001	<.001	<.001	<.001	<.001	<.001
Effect Size	0.11	0.52	0.6	0.47	0.58	0.56

Table E10. Identification of precise features for liver lesions. Median ICC (LCL-UCL).

		All [liver] – REPRO B	**All [liver] – REPRO R**	**All [liver] – Repeat R3B12**
First-order	**10Percentile**	**1.0 (1.0-1.0)**	**0.74 (0.73-0.76)**	**0.62 (0.6-0.64)**
	90Percentile	**1.0 (1.0-1.0)**	**0.74 (0.73-0.76)**	**0.56 (0.54-0.58)**
	Energy	**1.0 (1.0-1.0)**	**0.6 (0.58-0.62)**	**0.76 (0.75-0.78)**
	Entropy	0.97 (0.97-0.97)	0.44 (0.42-0.46)	0.45 (0.42-0.47)
	InterquartileRange	1.0 (1.0-1.0)	0.38 (0.36-0.4)	0.31 (0.28-0.34)
	Kurtosis	1.0 (1.0-1.0)	0.14 (0.11-0.16)	0.12 (0.1-0.15)
	Maximum	1.0 (1.0-1.0)	0.52 (0.5-0.55)	0.44 (0.41-0.46)
	Mean	**1.0 (1.0-1.0)**	**0.81 (0.8-0.82)**	**0.63 (0.6-0.64)**
	MeanAbsoluteDeviation	1.0 (1.0-1.0)	0.46 (0.44-0.49)	0.39 (0.36-0.41)
	Median	**1.0 (1.0-1.0)**	**0.79 (0.77-0.8)**	**0.61 (0.59-0.64)**
	Minimum	**1.0 (1.0-1.0)**	**0.57 (0.55-0.59)**	**0.55 (0.54-0.58)**
	Range	1.0 (1.0-1.0)	0.38 (0.35-0.4)	0.39 (0.36-0.42)
	RobustMeanAbsoluteDeviation	1.0 (1.0-1.0)	0.41 (0.39-0.43)	0.34 (0.31-0.37)
	RootMeanSquared	**1.0 (1.0-1.0)**	**0.82 (0.81-0.83)**	**0.62 (0.6-0.64)**
	Skewness	1.0 (1.0-1.0)	0.22 (0.2-0.24)	0.24 (0.21-0.26)
	Uniformity	0.94 (0.93-0.94)	0.4 (0.37-0.42)	0.43 (0.41-0.45)
	Variance	1.0 (1.0-1.0)	0.45 (0.43-0.48)	0.38 (0.35-0.41)
GLCM	**Autocorrelation**	**1.0 (1.0-1.0)**	**0.81 (0.8-0.82)**	**0.6 (0.58-0.62)**
	ClusterProminence	0.98 (0.98-0.98)	0.33 (0.3-0.36)	0.3 (0.27-0.33)
	ClusterShade	0.96 (0.96-0.96)	0.19 (0.16-0.22)	0.16 (0.12-0.19)
	ClusterTendency	0.98 (0.98-0.98)	0.42 (0.4-0.45)	0.35 (0.32-0.38)
	Contrast	0.95 (0.95-0.95)	0.49 (0.47-0.51)	0.41 (0.38-0.44)
	Correlation	0.97 (0.97-0.97)	0.27 (0.24-0.3)	0.36 (0.33-0.38)
	DifferenceAverage	0.93 (0.93-0.94)	0.51 (0.49-0.54)	0.43 (0.4-0.46)
	DifferenceEntropy	0.93 (0.92-0.93)	0.37 (0.34-0.4)	0.45 (0.43-0.47)
	DifferenceVariance	0.94 (0.93-0.94)	0.44 (0.42-0.46)	0.39 (0.37-0.42)
	Id	0.9 (0.89-0.9)	0.5 (0.47-0.52)	0.43 (0.41-0.46)
	Idm	0.9 (0.9-0.91)	0.5 (0.48-0.52)	0.44 (0.42-0.46)
	Idmn	0.95 (0.94-0.95)	0.5 (0.48-0.52)	0.42 (0.38-0.44)
	Idn	0.93 (0.92-0.93)	0.52 (0.5-0.54)	0.43 (0.4-0.46)
	Imc1	0.88 (0.87-0.88)	0.41 (0.38-0.44)	0.42 (0.39-0.45)
	Imc2	0.91 (0.91-0.92)	0.26 (0.23-0.28)	0.41 (0.39-0.44)
	InverseVariance	0.0 (0.0-0.0)	0.39 (0.36-0.41)	0.26 (0.24-0.29)
	JointAverage	**1.0 (1.0-1.0)**	**0.79 (0.78-0.8)**	**0.62 (0.6-0.64)**
	JointEnergy	0.88 (0.88-0.89)	0.41 (0.38-0.43)	0.45 (0.43-0.47)
	JointEntropy	0.94 (0.93-0.94)	0.48 (0.46-0.51)	0.5 (0.47-0.52)
	MaximumProbability	0.71 (0.69-0.72)	0.41 (0.39-0.43)	0.37 (0.35-0.4)
	SumAverage	**1.0 (1.0-1.0)**	**0.8 (0.79-0.82)**	**0.62 (0.6-0.64)**
	SumEntropy	0.97 (0.97-0.97)	0.43 (0.4-0.45)	0.45 (0.42-0.47)
	SumSquares	0.98 (0.98-0.98)	0.47 (0.44-0.49)	0.37 (0.34-0.4)
GLDM	**DependenceEntropy**	**0.86 (0.85-0.86)**	**0.56 (0.54-0.59)**	**0.64 (0.62-0.66)**
	DependenceNonUniformity	0.92 (0.91-0.92)	0.24 (0.21-0.27)	0.82 (0.81-0.83)
	DependenceNonUniformityNormalized	0.74 (0.72-0.75)	0.41 (0.39-0.44)	0.44 (0.42-0.46)
	DependenceVariance	0.66 (0.64-0.68)	0.41 (0.38-0.43)	0.33 (0.31-0.36)
	GrayLevelNonUniformity	**0.99 (0.99-0.99)**	**0.55 (0.53-0.57)**	**0.84 (0.83-0.85)**
	GrayLevelVariance	0.98 (0.98-0.98)	0.45 (0.42-0.47)	0.38 (0.35-0.4)
	HighGrayLevelEmphasis	**1.0 (1.0-1.0)**	**0.81 (0.8-0.82)**	**0.61 (0.59-0.63)**
	LargeDependenceEmphasis	0.84 (0.83-0.85)	0.49 (0.47-0.51)	0.52 (0.5-0.54)
	LargeDependenceHighGrayLevelEmphasis	0.86 (0.86-0.87)	0.51 (0.49-0.53)	0.5 (0.48-0.51)

	LargeDependenceLowGrayLevelEmphasis	**0.92 (0.92-0.93)**	**0.55 (0.53-0.57)**	**0.6 (0.58-0.62)**
	LowGrayLevelEmphasis	**0.97 (0.97-0.98)**	**0.53 (0.51-0.56)**	**0.59 (0.56-0.61)**
	SmallDependenceEmphasis	0.7 (0.68-0.71)	0.4 (0.38-0.43)	0.48 (0.46-0.51)
	SmallDependenceHighGrayLevelEmphasis	**0.76 (0.75-0.78)**	**0.55 (0.52-0.57)**	**0.53 (0.5-0.55)**
	SmallDependenceLowGrayLevelEmphasis	0.81 (0.8-0.82)	0.35 (0.32-0.37)	0.36 (0.33-0.39)
GLRLM	**GrayLevelNonUniformity**	**0.98 (0.98-0.98)**	**0.58 (0.56-0.6)**	**0.87 (0.86-0.87)**
	GrayLevelNonUniformityNormalized	0.95 (0.94-0.95)	0.39 (0.37-0.42)	0.44 (0.42-0.46)
	GrayLevelVariance	0.97 (0.97-0.98)	0.44 (0.41-0.46)	0.38 (0.35-0.41)
	HighGrayLevelRunEmphasis	**1.0 (1.0-1.0)**	**0.8 (0.79-0.81)**	**0.61 (0.59-0.63)**
	LongRunEmphasis	0.87 (0.86-0.88)	0.51 (0.49-0.54)	0.61 (0.59-0.62)
	LongRunHighGrayLevelEmphasis	**0.88 (0.87-0.88)**	**0.71 (0.69-0.73)**	**0.56 (0.54-0.58)**
	LongRunLowGrayLevelEmphasis	**0.92 (0.91-0.92)**	**0.55 (0.52-0.58)**	**0.62 (0.59-0.64)**
	LowGrayLevelRunEmphasis	**0.97 (0.97-0.97)**	**0.52 (0.5-0.55)**	**0.58 (0.56-0.61)**
	RunEntropy	0.9 (0.89-0.91)	0.44 (0.41-0.46)	0.51 (0.49-0.54)
	RunLengthNonUniformity	**0.95 (0.94-0.95)**	**0.68 (0.66-0.7)**	**0.95 (0.94-0.95)**
	RunLengthNonUniformityNormalized	0.88 (0.87-0.88)	0.5 (0.48-0.52)	0.54 (0.52-0.56)
	RunPercentage	**0.88 (0.88-0.89)**	**0.53 (0.51-0.55)**	**0.56 (0.54-0.58)**
	RunVariance	**0.84 (0.83-0.85)**	**0.54 (0.52-0.56)**	**0.6 (0.59-0.62)**
	ShortRunEmphasis	0.86 (0.85-0.87)	0.48 (0.46-0.5)	0.54 (0.52-0.56)
	ShortRunHighGrayLevelEmphasis	**0.98 (0.97-0.98)**	**0.8 (0.78-0.81)**	**0.62 (0.6-0.64)**
	ShortRunLowGrayLevelEmphasis	0.96 (0.96-0.96)	0.51 (0.49-0.54)	0.57 (0.55-0.6)
GLSZM	GrayLevelNonUniformity	0.52 (0.5-0.54)	0.15 (0.12-0.18)	0.34 (0.32-0.37)
	GrayLevelNonUniformityNormalized	0.58 (0.56-0.6)	0.27 (0.25-0.3)	0.29 (0.26-0.31)
	GrayLevelVariance	0.75 (0.74-0.77)	0.32 (0.3-0.35)	0.35 (0.32-0.38)
	HighGrayLevelZoneEmphasis	0.81 (0.8-0.82)	0.63 (0.61-0.65)	0.5 (0.47-0.53)
	LargeAreaEmphasis	0.82 (0.81-0.83)	0.4 (0.38-0.43)	0.68 (0.66-0.7)
	LargeAreaHighGrayLevelEmphasis	0.8 (0.79-0.81)	0.34 (0.32-0.37)	0.63 (0.61-0.65)
	LargeAreaLowGrayLevelEmphasis	0.86 (0.85-0.87)	0.5 (0.47-0.52)	0.68 (0.67-0.7)
	LowGrayLevelZoneEmphasis	0.86 (0.85-0.86)	0.4 (0.37-0.43)	0.46 (0.43-0.48)
	SizeZoneNonUniformity	0.38 (0.35-0.41)	0.11 (0.09-0.14)	0.24 (0.21-0.26)
	SizeZoneNonUniformityNormalized	0.2 (0.17-0.22)	0.06 (0.04-0.09)	0.09 (0.07-0.12)
	SmallAreaEmphasis	0.26 (0.23-0.28)	0.06 (0.04-0.09)	0.08 (0.06-0.11)
	SmallAreaHighGrayLevelEmphasis	0.31 (0.29-0.35)	0.25 (0.22-0.28)	0.21 (0.18-0.24)
	SmallAreaLowGrayLevelEmphasis	0.65 (0.63-0.66)	0.26 (0.23-0.28)	0.27 (0.23-0.3)
	ZoneEntropy	0.56 (0.54-0.58)	0.33 (0.3-0.36)	0.43 (0.4-0.45)
	ZonePercentage	0.72 (0.7-0.73)	0.48 (0.46-0.51)	0.57 (0.55-0.6)
	ZoneVariance	0.9 (0.89-0.9)	0.38 (0.36-0.41)	0.71 (0.7-0.73)
NGTDM	Busyness	0.73 (0.71-0.75)	0.04 (0.02-0.08)	0.48 (0.45-0.5)
	Coarseness	0.96 (0.95-0.96)	0.42 (0.38-0.44)	0.88 (0.88-0.89)
	Complexity	0.81 (0.8-0.82)	0.34 (0.32-0.37)	0.35 (0.32-0.38)
	Contrast	0.63 (0.61-0.65)	0.22 (0.2-0.26)	0.26 (0.23-0.28)
	Strength	0.87 (0.86-0.87)	0.35 (0.33-0.38)	0.42 (0.39-0.45)

Precise features identified if LCL ≥ 0.5 across the three variability sources.

Table E11. Identification of precise features for lung lesions (blue). Median ICC (LCL-UCL).

		All [lung]- Repro B	All [lung] – Repro R	All [lung] - Repeat
First-order	10Percentile	1.0 (1.0-1.0)	0.74 (0.71-0.76)	0.32 (0.28-0.38)
	90Percentile	**1.0 (1.0-1.0)**	**0.6 (0.57-0.64)**	**0.58 (0.54-0.62)**
	Energy	1.0 (1.0-1.0)	0.49 (0.44-0.53)	0.28 (0.24-0.31)
	Entropy	0.99 (0.99-0.99)	0.66 (0.63-0.7)	0.52 (0.48-0.55)
	InterquartileRange	1.0 (1.0-1.0)	0.7 (0.67-0.74)	0.34 (0.3-0.39)
	Kurtosis	1.0 (1.0-1.0)	0.24 (0.2-0.29)	0.19 (0.13-0.23)
	Maximum	1.0 (1.0-1.0)	0.48 (0.44-0.52)	0.63 (0.6-0.66)
	Mean	1.0 (1.0-1.0)	0.79 (0.77-0.81)	0.38 (0.34-0.42)
	MeanAbsoluteDeviation	1.0 (1.0-1.0)	0.75 (0.71-0.77)	0.37 (0.32-0.42)
	Median	1.0 (1.0-1.0)	0.73 (0.71-0.75)	0.35 (0.31-0.38)
	Minimum	1.0 (1.0-1.0)	0.53 (0.49-0.57)	0.35 (0.28-0.41)
	Range	1.0 (1.0-1.0)	0.55 (0.5-0.6)	0.36 (0.3-0.43)
	RobustMeanAbsoluteDeviation	1.0 (1.0-1.0)	0.73 (0.7-0.77)	0.35 (0.3-0.38)
	RootMeanSquared	1.0 (1.0-1.0)	0.71 (0.68-0.74)	0.37 (0.33-0.42)
	Skewness	1.0 (1.0-1.0)	0.31 (0.26-0.35)	0.21 (0.15-0.26)
	Uniformity	0.98 (0.98-0.98)	0.6 (0.57-0.64)	0.53 (0.49-0.58)
	Variance	1.0 (1.0-1.0)	0.66 (0.62-0.69)	0.29 (0.25-0.34)
GLCM	Autocorrelation	1.0 (1.0-1.0)	0.79 (0.78-0.81)	0.39 (0.36-0.43)
	ClusterProminence	1.0 (1.0-1.0)	0.41 (0.36-0.46)	0.13 (0.09-0.18)
	ClusterShade	1.0 (1.0-1.0)	0.29 (0.24-0.34)	0.09 (0.04-0.14)
	ClusterTendency	1.0 (1.0-1.0)	0.62 (0.59-0.66)	0.26 (0.21-0.31)
	Contrast	1.0 (1.0-1.0)	0.7 (0.67-0.73)	0.34 (0.3-0.38)
	Correlation	0.99 (0.99-0.99)	0.38 (0.34-0.42)	0.36 (0.31-0.41)
	DifferenceAverage	1.0 (1.0-1.0)	0.79 (0.77-0.82)	0.45 (0.41-0.49)
	DifferenceEntropy	0.99 (0.99-0.99)	0.44 (0.39-0.49)	0.48 (0.44-0.52)
	DifferenceVariance	1.0 (1.0-1.0)	0.5 (0.46-0.54)	0.24 (0.2-0.29)
	Id	**0.99 (0.99-0.99)**	**0.78 (0.76-0.81)**	**0.56 (0.52-0.61)**
	Idm	**0.99 (0.98-0.99)**	**0.77 (0.74-0.79)**	**0.57 (0.53-0.61)**
	Idmn	1.0 (1.0-1.0)	0.73 (0.7-0.75)	0.35 (0.32-0.39)
	Idn	1.0 (1.0-1.0)	0.81 (0.78-0.83)	0.47 (0.43-0.51)
	Imc1	**0.94 (0.94-0.95)**	**0.72 (0.68-0.74)**	**0.58 (0.54-0.61)**
	Imc2	0.96 (0.95-0.96)	0.55 (0.51-0.59)	0.48 (0.44-0.53)
	InverseVariance	**0.73 (0.7-0.75)**	**0.74 (0.7-0.76)**	**0.55 (0.5-0.59)**
	JointAverage	1.0 (1.0-1.0)	0.75 (0.73-0.78)	0.36 (0.32-0.4)
	JointEnergy	0.93 (0.92-0.94)	0.44 (0.4-0.48)	0.59 (0.55-0.62)
	JointEntropy	**0.92 (0.91-0.94)**	**0.55 (0.5-0.59)**	**0.62 (0.59-0.64)**
	MaximumProbability	0.85 (0.83-0.86)	0.45 (0.4-0.5)	0.51 (0.46-0.55)
	SumAverage	1.0 (1.0-1.0)	0.8 (0.78-0.81)	0.36 (0.32-0.4)
	SumEntropy	0.98 (0.98-0.98)	0.46 (0.4-0.5)	0.55 (0.51-0.58)
	SumSquares	1.0 (1.0-1.0)	0.66 (0.63-0.69)	0.28 (0.24-0.33)
GLDM	**DependenceEntropy**	**0.92 (0.91-0.92)**	**0.54 (0.5-0.58)**	**0.7 (0.66-0.73)**
	DependenceNonUniformity	0.94 (0.93-0.94)	0.2 (0.15-0.26)	0.46 (0.42-0.5)
	DependenceNonUniformityNormalized	**0.93 (0.92-0.94)**	**0.66 (0.63-0.7)**	**0.56 (0.52-0.59)**
	DependenceVariance	0.84 (0.83-0.86)	0.47 (0.42-0.51)	0.58 (0.56-0.61)
	GrayLevelNonUniformity	**1.0 (1.0-1.0)**	**0.77 (0.75-0.79)**	**0.76 (0.74-0.78)**
	GrayLevelVariance	1.0 (1.0-1.0)	0.65 (0.62-0.69)	0.29 (0.25-0.33)
	HighGrayLevelEmphasis	1.0 (1.0-1.0)	0.79 (0.77-0.8)	0.4 (0.37-0.44)
	LargeDependenceEmphasis	**0.95 (0.95-0.96)**	**0.62 (0.59-0.65)**	**0.64 (0.6-0.67)**

	LargeDependenceHighGrayLevelEmphasis	**0.95 (0.95-0.96)**	**0.63 (0.6-0.66)**	**0.62 (0.59-0.66)**
	LargeDependenceLowGrayLevelEmphasis	0.9 (0.89-0.91)	0.4 (0.35-0.46)	0.21 (0.16-0.26)
	LowGrayLevelEmphasis	0.95 (0.94-0.95)	0.37 (0.31-0.41)	0.02 (0.0-0.06)
	SmallDependenceEmphasis	**0.93 (0.93-0.94)**	**0.73 (0.71-0.76)**	**0.56 (0.52-0.6)**
	SmallDependenceHighGrayLevelEmphasis	**0.89 (0.88-0.9)**	**0.6 (0.57-0.64)**	**0.56 (0.52-0.6)**
	SmallDependenceLowGrayLevelEmphasis	0.96 (0.96-0.96)	0.36 (0.31-0.41)	0.05 (0.01-0.08)
GLRLM	**GrayLevelNonUniformity**	**0.99 (0.99-0.99)**	**0.8 (0.78-0.82)**	**0.76 (0.74-0.79)**
	GrayLevelNonUniformityNormalized	0.98 (0.98-0.99)	0.61 (0.58-0.65)	0.52 (0.49-0.57)
	GrayLevelVariance	1.0 (1.0-1.0)	0.65 (0.61-0.68)	0.29 (0.25-0.33)
	HighGrayLevelRunEmphasis	1.0 (1.0-1.0)	0.78 (0.76-0.8)	0.41 (0.38-0.45)
	LongRunEmphasis	**0.96 (0.96-0.97)**	**0.7 (0.68-0.72)**	**0.62 (0.59-0.65)**
	LongRunHighGrayLevelEmphasis	**0.96 (0.96-0.96)**	**0.76 (0.74-0.78)**	**0.58 (0.54-0.62)**
	LongRunLowGrayLevelEmphasis	0.86 (0.85-0.88)	0.35 (0.29-0.4)	0.01 (0.0-0.06)
	LowGrayLevelRunEmphasis	0.95 (0.95-0.96)	0.36 (0.3-0.4)	0.03 (0.0-0.06)
	RunEntropy	0.97 (0.97-0.97)	0.61 (0.57-0.66)	0.49 (0.44-0.53)
	RunLengthNonUniformity	**0.93 (0.92-0.94)**	**0.63 (0.59-0.66)**	**0.92 (0.91-0.93)**
	RunLengthNonUniformityNormalized	**0.97 (0.97-0.97)**	**0.73 (0.7-0.76)**	**0.62 (0.59-0.65)**
	RunPercentage	**0.97 (0.97-0.98)**	**0.73 (0.71-0.76)**	**0.63 (0.6-0.66)**
	RunVariance	**0.95 (0.94-0.95)**	**0.72 (0.7-0.75)**	**0.63 (0.6-0.66)**
	ShortRunEmphasis	**0.97 (0.97-0.97)**	**0.7 (0.68-0.73)**	**0.62 (0.58-0.65)**
	ShortRunHighGrayLevelEmphasis	0.95 (0.94-0.95)	0.73 (0.72-0.75)	0.39 (0.33-0.43)
	ShortRunLowGrayLevelEmphasis	0.95 (0.94-0.95)	0.36 (0.3-0.4)	0.04 (0.0-0.07)
GLSZM	GrayLevelNonUniformity	0.59 (0.56-0.62)	0.27 (0.21-0.33)	0.54 (0.51-0.58)
	GrayLevelNonUniformityNormalized	0.9 (0.89-0.91)	0.54 (0.49-0.58)	0.44 (0.41-0.49)
	GrayLevelVariance	0.97 (0.97-0.97)	0.55 (0.51-0.6)	0.3 (0.24-0.34)
	HighGrayLevelZoneEmphasis	0.95 (0.94-0.95)	0.62 (0.59-0.65)	0.44 (0.39-0.47)
	LargeAreaEmphasis	**0.94 (0.93-0.94)**	**0.6 (0.57-0.64)**	**0.6 (0.56-0.63)**
	LargeAreaHighGrayLevelEmphasis	**0.94 (0.93-0.94)**	**0.61 (0.57-0.64)**	**0.6 (0.56-0.63)**
	LargeAreaLowGrayLevelEmphasis	0.91 (0.91-0.92)	0.43 (0.36-0.48)	0.51 (0.47-0.56)
	LowGrayLevelZoneEmphasis	0.95 (0.94-0.96)	0.28 (0.22-0.33)	0.03 (0.0-0.06)
	SizeZoneNonUniformity	0.81 (0.79-0.83)	0.41 (0.36-0.46)	0.46 (0.41-0.49)
	SizeZoneNonUniformityNormalized	0.62 (0.58-0.65)	0.46 (0.41-0.51)	0.4 (0.34-0.44)
	SmallAreaEmphasis	0.65 (0.61-0.69)	0.47 (0.42-0.52)	0.39 (0.34-0.43)
	SmallAreaHighGrayLevelEmphasis	0.55 (0.52-0.58)	0.31 (0.26-0.36)	0.22 (0.18-0.27)
	SmallAreaLowGrayLevelEmphasis	0.95 (0.94-0.95)	0.27 (0.22-0.32)	0.04 (0.0-0.07)
	ZoneEntropy	0.84 (0.82-0.85)	0.47 (0.42-0.52)	0.51 (0.48-0.56)
	ZonePercentage	**0.94 (0.93-0.94)**	**0.76 (0.74-0.79)**	**0.58 (0.55-0.61)**
	ZoneVariance	**0.95 (0.94-0.96)**	**0.55 (0.52-0.6)**	**0.63 (0.6-0.66)**
NGTDM	Busyness	0.92 (0.91-0.92)	0.04 (0.0-0.08)	0.43 (0.39-0.48)
	Coarseness	0.98 (0.97-0.98)	0.0 (0.0-0.0)	0.81 (0.79-0.83)
	Complexity	0.98 (0.97-0.98)	0.6 (0.56-0.64)	0.3 (0.27-0.33)
	Contrast	0.94 (0.93-0.95)	0.54 (0.5-0.58)	0.42 (0.38-0.46)
	Strength	0.97 (0.97-0.98)	0.63 (0.59-0.68)	0.34 (0.29-0.38)

Precise features identified if LCL ≥ 0.5 across the three variability sources.

Table E12. Lists of precise and non-precise RF selected for habitat computation after redundancy elimination (i.e., Spearman r ≤ 0.7) for the liver lesion displayed in Figure 5.

Precise RF	Non-Precise RF
FO_Kurtosis	GLDM_DependenceEntropy
GLCM_ClusterShade	GLRLM_LowGrayLevelRunEmphasis
GLRLM_ GrayLevelNonUniformityNormalized	GLRLM_RunLengthNonUniformity
GLSZM _GrayLevelNonUniformity	GLRLM_RunPercentage
GLSZM_SmallAreaEmphasis	GLRLM_RunVariance
GLSZM_SmallAreaHighGrayLevelEmphasis	NGTDM_Coarseness
GLSZM_SmallAreaLowGrayLevelEmphasis	
GLSZM_ZoneEntropy	
GLSZM_ZonePercentage	
NGTDM_Busyness	
NGTDM_Complexity	
NGTDM _Contrast	
NGTDM_Strength	

FO, first-order; GLCM**,** Grey Level Co-occurrence Matrix features; GLDM**,** Grey Level Dependence Matrix; GLRLM**,** Grey Level Run Length Matrix; GLSZM**,** Grey Level Size Zone Matrix; NGTDM**,** Neighbouring Grey Tone Difference Matrix Features.

FIGURE LEGENDS

FIGURE 1. A) Distribution of lung and liver lesions across different cohorts for precision analysis. B) Precision analysis design. Three-dimensional (3D) radiomics features were computed from both original and perturbed images, four times per image, each time with a different combination of kernel radius, R (1mm/3mm), and bin size, B (12 Hounsfield units [HU]/25HU). To study repeatability, original-perturbed feature pairs were evaluated for every combination of computation settings (R1B12, R1B25, R3B12 and R3B25). To study reproducibility against computation parameters, we compared pairs of original features computed under varying settings. To understand reproducibility against the kernel radius (R), we kept the bin size (B) constant at two separate levels (12HU and 25HU) and then altered the kernel radius. To explore reproducibility against the bin size (B), we kept the kernel radius (R) constant at two different measures (1mm and 3mm) and varied the bin size. Precise features were selected by linking reproducibility and repeatability results.

FIGURE 2. A) Repeatability distribution of radiomics features per setting. Most radiomics features exhibited poor repeatability. Features computed with kernel radius (R) of 3mm were more repeatable than those computed with R=1mm. Bin size changes did not affect repeatability. B) Repeatability distribution of radiomics features computed with setting R3B12 per feature class for lung and liver lesions separately. First order and GLRLM features were more repeatable in liver lesions while GLCM features were more repeatable in lung lesions. LCL, 95% lower confidence limit of the intraclass correlation coefficient; R3B12, features computed with kernel radius 3mm and bin size 12 Hounsfield units; FO, First-Order; GLCM, Grey Level Co-occurrence Matrix features; GLDM, Grey Level Dependence Matrix; GLRLM, Grey Level Run Length Matrix; GLSZM, Grey Level Size Zone Matrix; NGTDM, Neighboring Grey Tone Difference Matrix Features.

FIGURE 3. A) Reproducibility distribution against kernel radius, R, for features computed with fixed bin size of 12 Hounsfield units (HU) and bin size 25HU. Most features showed poor reproducibility against R. Features computed with B=12HU were more reproducible ($p<.001$). B) Reproducibility distribution against bin size, B, for features

computed with fixed kernel radius 3mm and fixed kernel radius 1mm. Most features showed good or excellent reproducibility against B. Features computed with R=3mm were more reproducible (p<.001). C) Reproducibility distribution against kernel radius for features computed with fixed bin size of 12HU per feature class for lung and liver lesions separately. Features computed from lung lesions were more reproducible against R, especially for features belonging to GLCM and GLRLM classes. D) Reproducibility distribution against bin size for features computed with fixed kernel radius 3mm per feature class for lung and liver lesions separately. Features computed from lung lesions are more reproducible against B, especially for features belonging to GLCM, first-order and NGTM classes. LCL, 95% lower confidence limit of the intraclass correlation coefficient; FO, First-Order features; GLCM, Grey Level Co-occurrence Matrix features; GLDM, Grey Level Dependence Matrix; GLRM, Grey Level Run Length Matrix; GSZM, Grey Level Size Zone Matrix; NGTM, Neighboring Grey Tone Difference Matrix Features.

FIGURE 4. Heatmap displaying the lower 95% confidence limit (LCL) of the intraclass correlation coefficient results obtained in the three experiments used to identify precise features: repeatability (setting R3B12), reproducibility against R (fixed B=12 Hounsfield units), and reproducibility against B (fixed R=3mm), for lung and liver lesions separately. FO, First-Order features; GLCM, Grey Level Co-occurrence Matrix features; GLDM, Grey Level Dependence Matrix; GLRLM, Grey Level Run Length Matrix; GLSZM, Grey Level Size Zone Matrix; NGTDM, Neighboring Grey Tone Difference Matrix Features.

FIGURE 5. A) Original and perturbed CT scans for one liver lesion (84-year-old male). B) Example of habitats obtained for the same lesion. Habitats computed with precise features show higher stability (measured via Dice similarity coefficient [DSC] of original-perturbed habitat pairs). Top row: habitats obtained with precise features computed from original image (left) and perturbed image (right). DSC scores for habitats 1, 2 and 3 are 0.976, 0.891 and 0.915, respectively. Bottom row: habitats obtained with non-precise (i.e., all computed features) features computed from original image (left) and perturbed image (right). DSC scores for habitats 1, 2 and 3 are 0.751, 0.328 and 0.57, respectively. C) Quantification of habitat stability computed with precise features and non-precise features for all lung and liver lesions. Boxes represents the IQR (25th–75th percentile), and the horizontal line inside the boxes represents the median value of the DSC. Whiskers represent the minimum and maximum values, ♦ = outliers, ****=

$p < 0.0001$. Habitats computed with precise features show higher stability (Wilcoxon signed rank test, $p < 0.0001$)..

SUPPLEMENTAL FIGURE LEGENDS

FIGURE E1. Example of liver lesion (original and perturbed) and axial section of features First-Order Kurtosis and NGTDM Strength from original and perturbed images computed with 4 settings. R1B12, features computed with kernel radius 1mm and bin size 12HU; R1B25, features computed with kernel radius 1mm and bin size 25HU; R3B12, features computed with kernel radius 3mm and bin size 12HU; R3B25, features computed with kernel radius 3mm and bin size 25HU; NGTDM, Neighboring Grey Tone Difference Matrix Features; Ori, original; Pert, perturbed.

FIGURE E2. Optimal number of habitats selection. The graph displays the gradient of the Bayesian Information Criterion (BIC) against the number of clusters (k) or habitats. When fitting the data (i.e. clustering the voxel-wise features), BIC penalizes the addition of parameters that result in overfitting[30]. As it is observed, the larger the number of clusters, the larger the gradient, meaning the original BIC function keeps decreasing (and thus the likelihood increasing). However, starting from a cluster size of three the gradient of the BIC increases slower, i.e. the original function has a gentler decrease. Thus, increasing the number of clusters beyond k=3 does not result in additional information gain.

FIGURE E3. Repeatability distribution of radiomics features computed with setting R1B12 (A), R1B25 (B), R3B12 (C) and R3B25 (D) per cohort for lung and liver lesions separately. Primary tumor has no effect on repeatability. LCL, 95% lower confidence limit of the Intraclass Correlation Coefficient; R1B12, features computed with kernel radius 1mm and bin size 12HU; R1B25, features computed with kernel radius 1mm and bin size 25HU; R3B12, features computed with kernel radius 3mm and bin size 12HU; R3B25, features computed with kernel radius 3mm and bin size 25HU; CRC: colorectal cohort; NET: neuroendocrine cohort; ALL: all cohorts combined.

FIGURE E4. Heatmap displaying results obtained in the four repeatability experiments (one per setting) for lung and liver lesions separately. LCL, 95% lower confidence limit of the Intraclass Correlation Coefficient; R1B12, features computed with kernel radius 1mm and bin size 12HU; R1B25, features computed with kernel radius 1mm and bin size 25HU; R3B12, features computed with kernel radius 3mm and bin size 12HU; R3B25, features computed with kernel radius 3mm and bin size 25HU; FO, First-Order; GLCM, Grey Level Co-occurrence Matrix features; GLDM, Grey Level Dependence Matrix; GLRLM, Grey Level Run Length Matrix; GLSZM, Grey Level Size Zone Matrix; NGTDM, Neighboring Grey Tone Difference Matrix Features.

FIGURE E5. Reproducibility distribution against R of radiomics features computed with fixed bin size of 12HU (A) and fixed bin size of 25HU (B) per cohort for lung and liver lesions separately. Similary, (C) and (D) depict the reproducibility distribution against B of radiomics features computed with fixed radius of 1mm (C) and 3mm (D) per cohort for lung and liver lesions separately. LCL, 95% lower confidence limit of the Intraclass Correlation Coefficient; CRC: colorectal cohort; NET: neuroendocrine cohort; ALL: all cohorts combined.

FIGURE E6. Heatmap displaying results obtained in the four reproducibility experiments for lung and liver lesions separately: reproducibility against R (fixed B=12HU), reproducibility against R (fixed B=25HU), reproducibility against B (fixed R=1mm), reproducibility against B (fixed R=3mm). LCL, 95% lower confidence limit of the Intraclass Correlation Coefficient; FO, First-Order; GLCM, Grey Level Co-occurrence Matrix features; GLDM, Grey Level Dependence Matrix; GLRLM, Grey Level Run Length Matrix; GLSZM, Grey Level Size Zone Matrix; NGTDM, Neighboring Grey Tone Difference Matrix Features.

FIGURE E7. Correlation heatmap of precise RF prior to habitat computation for one liver lesion (Figure 5). Correlation was assessed using Spearman rank correlation coefficient, with p<.005.

FIGURE E8. Correlation heatmap of non-precise RF prior to habitat computation for one liver lesion (Figure 5).

Correlation was assessed using Spearman rank correlation coefficient, with p<.005.

FIGURE E9. Number of selected habitats by the Bayesian Information Criterion (BIC) for both CT and mpMRI habitats in the independent cohort (13 liver lesions). Each circle represents a lesion. Orange circles represent 2 computed habitats and green circles 3 computed habitats. The model computed the same number of habitats for both modalities in 10 out of 13 patients, indicating a that the algorithm is robust in classifying regions with different phenotypes with both CT and mpMRI data.

FIGURE E10. Exploration of the biological relevance of imaging habitats in an independent cohort. One representative patient (liver metastasis of melanoma). A) CT scan with visible lesion (yellow arrow) and resulting CT habitats computed with precise liver radiomics features (also shown). B) Anatomical T2-weighted MRI scan with visible lesion (yellow arrow) and resulting multiparametric MRI (mpMRI) habitats computed with the 9 mpMRI maps (also shown). mpMRI and CT habitats showed comparable distributions. C) Image-guided biopsy with needle (N) and liver (L) tumor lesion (T) and resulting histologic image (hematoxylin and eosin [HE] stain [400x magnification]) with observable tissue types, annotated by a pathologist. The HE-stained histologic material reveals areas of necrosis in the core of the lesion. GDM: Gray Level Dependence Matrix, GDM_DE: Dependence Entropy, GLRLM: Gray Level Run Length Matrix, GRLM_GNU: Gray Level Non Uniformity, GRLM_LGRE: Low Gray Level Run Emphasis, GRLM_RLNU: Run Length Non Uniformity, GRLM_SRHGE: Short Run High Gray Level Emphasis, GRLM_RV: Run Variance, GRLM_RP: Run Percentage, NTGDM: Neighboring Gray Tone Difference Matrix, NGTDM_CO: Coarseness. ADCt: tissue apparent diffusion coefficient, ADCv: vascular apparent diffusion coefficient, T2t: tissue transvers relaxation time, AKt: tissue apparent kurtosis coefficient, Fv; vascular signal fraction, T1: total longitudinal relaxation time, KTrans: capillary permeability constant, Vp: plasma volume fraction, Ve: extravascular and extracellular volume fraction.

FIGURE E11. Exploration of the biological relevance of imaging habitats (case study). Representative patient (liver metastasis). A) CT scan with visible lesion (yellow arrow) and resulting CT habitats computed with precise liver radiomic features. B) Anatomical MRI T2 scan with visible lesion (yellow arrow) and resulting mpMRI habitats. C)

Histologic image (HE stain [400x magnification]) with observable tissue types, annotated by a pathologist. The HE-stained histologic material reveals areas of fibrosis within the lesion that may correspond to the yellow habitat depicted in A and B.