Radiomics in Oncology: A Practical Guide

Joshua D. Shur, MBBS
Simon J. Doran, PhD
Santosh Kumar, PhD
Derfel ap Dafydd, MBBS
Kate Downey, MD
James P. B. O'Connor, PhD
Nikolaos Papanikolaou, PhD
Christina Messiou, MD
Dow-Mu Koh, MD
Matthew R. Orton, PhD

Radiomics refers to the extraction of mineable data from medical imaging and has been applied within oncology to improve diagnosis, prognostication, and clinical decision support, with the goal of delivering precision medicine. The authors provide a practical approach for successfully implementing a radiomic workflow from planning and conceptualization through manuscript writing. Applications in oncology typically are either classification tasks that involve computing the probability of a sample belonging to a category, such as benign versus malignant, or prediction of clinical events with a time-to-event analysis, such as overall survival. The radiomic workflow is multidisciplinary, involving radiologists and data and imaging scientists, and follows a stepwise process involving tumor segmentation, image preprocessing, feature extraction, model development, and validation. Images are curated and processed before segmentation, which can be performed on tumors, tumor subregions, or peritumoral zones. Extracted features typically describe the distribution of signal intensities and spatial relationship of pixels within a region of interest. To improve model performance and reduce overfitting, redundant and nonreproducible features are removed. Validation is essential to estimate model performance in new data and can be performed iteratively on samples of the dataset (cross-validation) or on a separate hold-out dataset by using internal or external data. A variety of noncommercial and commercial radiomic software applications can be used. Guidelines and artificial intelligence checklists are useful when planning and writing up radiomic studies. Although interest in the field continues to grow, radiologists should be familiar with potential pitfalls to ensure that meaningful conclusions can be drawn.

Online supplemental material is available for this article.

Abbreviations: GLCM = gray-level co-occurrence matrix, IBSI = Image Biomarker Standardization Initiative, ICC = intra-class correlation coefficient, ROI = region of interest, SI = signal intensity

RadioGraphics 2021; 41:1717–1732

https://doi.org/10.1148/rg.2021210037

Content Codes: AI OI

From the Department of Radiology, Royal Marsden Hospital NHS Foundation Trust, Sutton, England (J.D.S., D.a.D., K.D., N. P., C.M., D.M.K.); Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, England (S.J.D., S.K., J.P.B.O., N. P., C.M., D.M.K., M.R.O.); and Computational Clinical Imaging Group, Champalimaud Foundation, Centre for the Unknown, Lisbon, Portugal (N.P.). Received February 15, 2021; revision requested June 15 and received June 27; accepted June 29. For this journal-based SA-CME activity, the authors S.J.D., N.P., and D.M.K. have provided disclosures (see end of article); all other authors, the editor, and the reviewers have disclosed no relevant relationships. **Address correspondence to** M.R.O. (e-mail: *matthew.orton@icr.ac.uk*).

This article represents independent research funded by the National Institute for Health Research (NIHR) Biomedical Research Centre and Clinical Research Facility at The Royal Marsden NHS Foundation Trust and Institute of Cancer Research, London.

SA-CME LEARNING OBJECTIVES

After completing this journal-based SA-CME activity, participants will be able to:

- List the main applications of radiomic studies in oncology.
- Understand the use of image preprocessing, segmentation, and validation in radiomic studies.
- Describe the main radiomic feature classes and how they are calculated.

See www.rsna.org/education/search/RG.

Introduction

Radiomics refers to the extraction of mineable high-dimensional data from radiologic images (1–3) and has been applied within oncology to improve diagnosis and prognostication (4,5) with the aim of delivering precision medicine. The premise is that imaging data convey meaningful information about tumor biology, behavior, and pathophysiology (6) and may reveal information that is not otherwise apparent to current radiologic and clinical interpretation.

The radiomic workflow involves curation of clinical and imaging data and is a stepwise process involving image preprocessing, tumor segmentation, feature extraction, model development, and validation (7). It is a field that requires input from individuals in many disciplines, including radiologists, imaging scientists, and data scientists. Features are derived at single (usually pretreatment) or multiple (eg, δ radiomics) time points and can be applied to the whole spectrum of imaging data.

TEACHING POINTS

- Radiomics refers to the extraction of mineable high-dimensional data from radiologic images and has been applied within oncology to improve diagnosis and prognostication with the aim of delivering precision medicine.
- Radiomic studies in oncology are usually either *(a)* classification tasks or *(b)* prediction of clinical outcomes by using a time-to-event analysis.
- As with any research study, a radiomic study should have a testable hypothesis that should address a relevant clinical question, usually with the aim of meeting an unfulfilled need in cancer management.
- Radiomic features are "handcrafted" in that the algorithms used to generate them are designed or chosen by the data scientist rather than being learned directly from the images, as is found with deep learning approaches.
- To aid authors and to provide a framework for manuscript writing, there are various radiomic- and artifical intelligence–specific checklists, reporting guides, and radiomic quality scores that can be referred to, in addition to artifical intelligence extensions of familiar guidelines such as TRIPOD, CONSORT, and SPIRIT.

Although many of the concepts of image feature extraction have been around for decades (8), research output in the field has grown exponentially, with over 1500 publications in 2020 containing the term *radiomics* (Fig 1). With increasing interest in the field, there is a need for an understanding of the radiomic workflow and its challenges and limitations so that robust conclusions may be drawn (9). The purpose of this article is to provide a practical hands-on guide for implementing radiomic studies in oncology and a glossary of terms for readers less familiar with the topic (Table 1).

Applications in Oncology

Radiomic studies in oncology are usually either *(a)* classification tasks or *(b)* prediction of clinical outcomes by using a time-to-event analysis. Classification involves dividing a population into categories. Examples include benign versus malignant, genomic status, tumor stage, and presence of metastases, among many others. Predictive models use clinical outcomes to stratify patients into different risk groups on the basis of the risk of occurrence of clinical endpoints, such as overall or disease-free survival, and are assessed by using a time-to-event analysis.

These applications are guided by the notion that radiomic data convey information about tumor biology (1). For example, radiomic features may reflect temporal and spatial heterogeneity (Fig 2), which is known to be a key determinant of tumor behavior and resistance to therapy (10). Thus, radiomics has the potential to act as a "virtual biopsy" and, unlike standard biopsies, uses noninvasive imaging that permits analysis of the whole tumor (rather than a focal sample) and can be applied more easily at multiple time points for disease monitoring, offering potentially important diagnostic information related to disease evolution.

Planning a Radiomic Study

When planning a radiomic study, it is worth asking basic questions (Table 2) to assess feasibility and likelihood of success. At our institution, we find a radiomic study proforma useful when assessing proposed studies (Appendix E1). As with any research study, a radiomic study should have a testable hypothesis that should address a relevant clinical question, usually with the aim of meeting an unfulfilled need in cancer management.

A key consideration is to determine the availability of sufficient data to support the development of a *radiomic signature* (defined as the learned model from a radiomic analysis used to predict a particular clinical outcome). As a rule of thumb for binary classification studies, one should aim to obtain 10–15 samples per feature in the final radiomic signature. This can vary between studies but is a useful guide when embarking on a new study (11,12). If the class sizes are unequal, the rule should be applied to the smaller class (13). As radiomics is data driven, it may not be possible to know in advance how many features will be included in the final model, since feature selection methods are typically applied before or during the model fitting process. It is also important to be aware that data attrition is common. Common reasons include data that are missing or mislabeled, failure to satisfy inclusion criteria or lost to follow-up, and poor image quality. These highlight the importance of obtaining a realistic estimate of the final sample size before embarking on a new study.

Model validation consists of measuring the predictive performance of the model by using data that were not used in fitting the model. Sufficient data should be available for validation of a radiomic model, typically around one-third of the training sample size. The one-third proportion represents a trade-off between having enough data in the training set to ensure the model has sufficient predictive power and having a large-enough test dataset to ensure the predicted performance estimate is accurate. Values used in practice are in the range of 60:40 to 90:10. For example, using the "one-third" criteria and a 10-feature model, at least 133 samples are required, where 100 are used for training and 33 for validation. Assuming an attrition rate of 50% would require a total study population of 266. This highlights the challenge required to

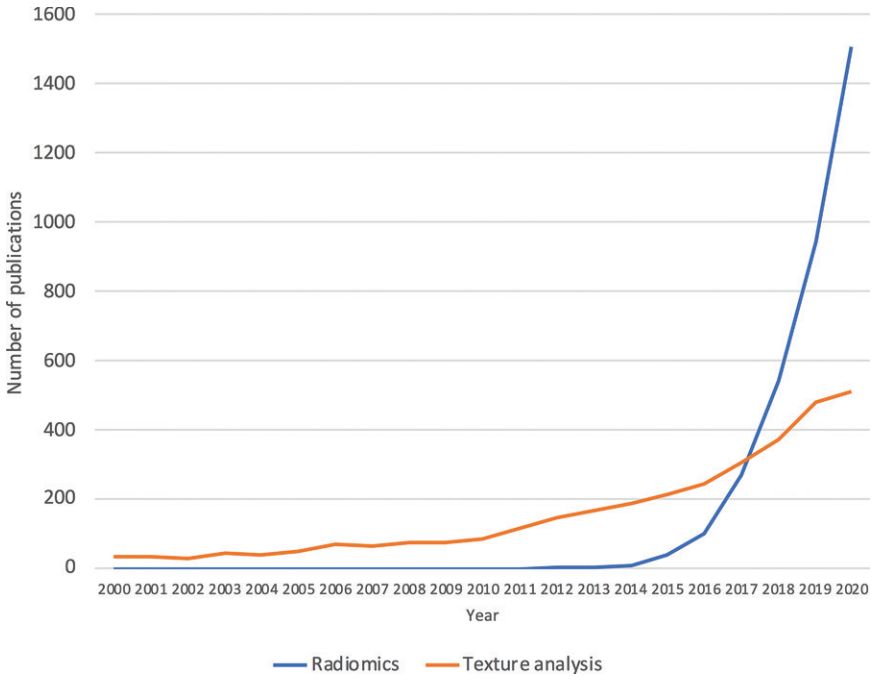


Figure 1. Graph shows the number of publications per year since 2000 that contain the terms *radiomics* and *texture analysis* in PubMed *(www.pubmed.gov)*. Since first being coined in 2012, the term *radiomics* in the literature has demonstrated an exponential increase, numbering over 1500 publications in 2020 alone. The term *radiomics* has overtaken *texture analysis* in publications in PubMed, indicating a shift toward *radiomics* as the preferred term in the research literature.

curate datasets of sufficient size for high-quality radiomic studies.

Finally, it is important to consider whether the data are balanced. For classification tasks, balanced data are such that each class or outcome contains an approximately equal proportion of the data. When proportions are unequal, the data are unbalanced, and if they are very unbalanced then a larger sample size may be required for the model developed to be generalizable. For time-to-event analyses, the proportion of events that are observed (event time known) and censored (eg, subject leaves the study before the event occurs or the study ends before the event occurs) should be estimated.

Consideration should be given to data heterogeneity, including disease status, treatment, imaging equipment, acquisition protocol, and method of measurement. There exists a trade-off between real-world heterogeneous datasets, in which noise may mask an underlying radiomic signature, and well-controlled homogeneous datasets that are less noisy but have lower generalizability. An assessment of data heterogeneity is performed by evaluating how similar the study design and inclusion criteria are compared with what is encountered in clinical practice. This will assist in assessing both the chance of success and also whether a follow-up real-world study would be required to establish a clinically useful signature.

Once the research question and study population have been defined, one should consider collecting pilot data to help identify and mitigate potential problems before full data collection. With a representative sample of data, frequency of missing data and rates of passing inclusion criteria can be estimated. For classification studies, a pilot sample size of 12 per class has been proposed (14), but in practice this is guided by available resources and the study population. Running pilot data through the radiomic processing pipeline as early as possible enables issues to be resolved quickly, and any preliminary results may be able to guide the final sample size. For example, if a signature is detected but is not statistically significant, then it may be possible to estimate the number of samples required to obtain a significant result.

Radiomic Workflow Overview

The radiomic workflow represents the combined effort of a multidisciplinary team, including data and imaging scientists and radiologists, and is subdivided into multiple tasks that are typically performed in sequence (Figs 3, 4).

Image Acquisition

Although the majority of studies to date have used data from CT examinations, radiomic analyses can be applied to the whole spectrum of imaging data, including those from CT, PET, MRI, and US examinations. One advantage of CT and PET data are that signal intensities (SIs) are inherently quantitative. CT may also be less prone to motion artifacts seen with PET and MRI. US is more user dependent than other modalities; however, along with MRI, assessing feature stability in a test-retest experiment is feasible, as there is no radiation burden. Ultimately the choice of modality is often determined by what is available and used in clinical practice.

Table 1: Glossary of Radiomics Terms

Term	Definition
Algorithm	See "model"
Balanced data	Each class or outcome contains an approximately equal proportion of data or number of samples. When proportions in each class are unequal, then the data are unbalanced.
Censoring	A patient is censored in a time-to-event analysis when the time-to-event is not known because of missing data (eg, patient is lost to follow-up).
Classification	Classifying samples into groups or categories on the basis of a classification rule; *binary* classification refers to two classes
δ-radiomics	Characterizing the change (δ) in feature values by applying radiomics to multiple time points (eg, before and after treatment)
Gaussian distribution	Also known as the normal distribution
Heterogeneity (data)	Differences in the data between patients, which may include aspects of the disease, treatment, imaging equipment used, acquisition protocol performed, or method of measurement used
Input feature	Features that are used to train the model; these typically include radiomic features and clinical (nonradiomic) features
Labeled data	Data that have been tagged with a label or class; for example, in a classification task, this label may specify the type of lesion (eg, cyst, hemangioma, metastasis)
Metadata	Data that describe or give information about other data; examples with imaging data include date of acquisition or acquisition parameter
Model	A mathematical function that can be used to predict the target features from the input features. Models have parameters, and the values of those parameters must be estimated from the training data by using a learning algorithm. The learned model can then be used on test data for validation or deployed on new data for "live" prediction.
Multivariate	Involving multiple variables or features
Observed	In time-to-event analyses, these are the events that occur.
Overfitting	The model performs well on the training data but poorly on the unseen validation data; this may occur if the number of features is large compared with the sample size and therefore captures "noise."
Radiomic signature	The learned model from a radiomic analysis used to predict a particular clinical outcome
Recursive	In computer science, recursive refers to defining a problem in terms of itself and involves repeatedly applying the same updating rule to something, usually with another rule for when to stop.
Redundancy	Refers to features that do not add any additional information to the input data. Two features may be redundant if they are highly correlated with one another, so excluding one feature will not impact the prediction performance.
Regularization	In machine learning, regularization is a technique for reducing the importance of some features within the statistical model to prevent overfitting.
Rule of 10	Widely used rule of thumb that suggests a minimum of 10 samples per feature included in the final radiomic model
Sample	A data point is referred to as a sample. In radiomics, this usually refers to the patient.
Supervised learning	Supervised learning uses labeled datasets (eg, benign vs malignant) to train computational models that classify data or predict outcomes on unseen data.
Target feature/data	The data that the model is trying to predict (eg, benign vs malignant)
Training dataset	The dataset used to train the learning algorithm or statistical model
Tuning parameters	Model parameters that affect the model behavior but for which the value cannot be estimated from a single training set (number of selected features, amount of regularization, tree depth, etc)
Underfitting	The learned model fails to capture certain patterns in the input data that are informative, leading to suboptimal performance; usually due to insufficient features within the model.
Univariate	Involving a single variable or feature
Unsupervised learning	As opposed to supervised learning, unsupervised learning refers to the use of algorithms to learn patterns in unlabeled datasets.
Validation dataset	The dataset used to measure the performance of the learning algorithm or statistical model. With hold-out validation, the training and validation datasets are typically split between a 60:40 and 90:10 ratio. Often one-third is used as a rule of thumb.

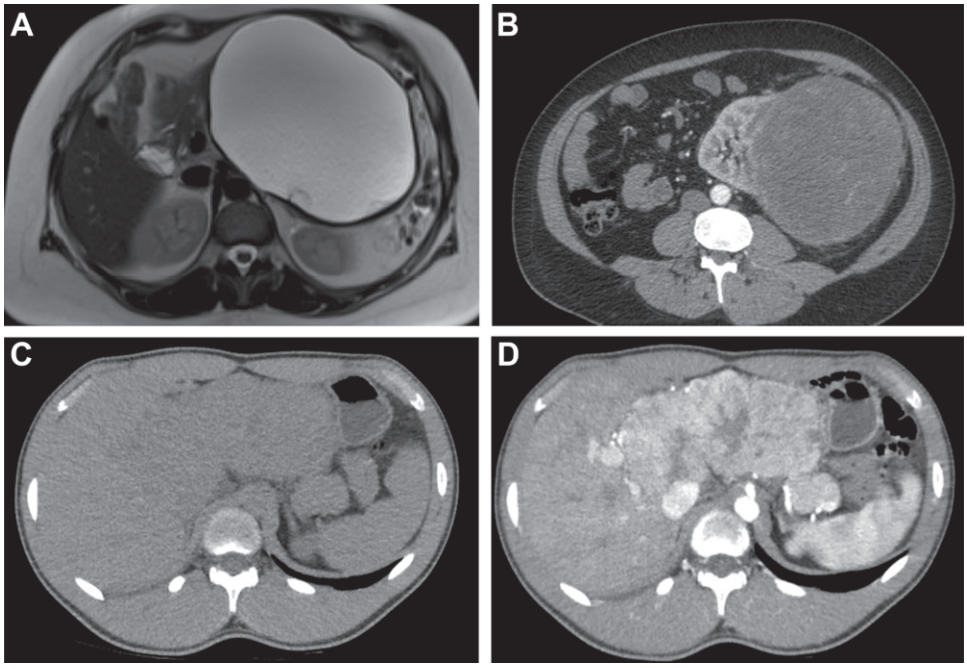


Figure 2. Variations in tumor heterogeneity from less to more heterogeneous are demonstrated in these abdominal masses. **(A)** Axial T2-weighted MR image in a 40-year-old woman shows a large unilocular cystic lesion in the pancreas that appears to have uniform high signal intensity (SI), with only minor nonenhancing peripheral septa and a smooth border. This appearance is typical for a mucinous cystadenoma. After surgical resection, no invasive malignancy was found. **(B)** Axial CT image shows a partly heterogeneous mass in the left kidney, which appears well defined and contains predominantly homogeneous bland-appearing tissue with streaks of vascularity. This was found to be a spindle-cell sarcoma after surgical resection. **(C, D)** Axial nonenhanced **(C)** and contrast-enhanced **(D)** CT images of a fibrolamellar hepatocellular carcinoma clearly show the heterogeneous nature of this malignant tumor, with irregular vascular enhancing tissue surrounding a less-vascular central component. Contrast-enhanced imaging is often used in radiomic analyses and is useful to help highlight vascularity and spatial heterogeneity, a determinant of tumor behavior and resistance to therapy that is not readily apparent without contrast material.

Table 2: Questions for Planning a Radiomic Study

Question	Action
What is the likely impact of the proposed study?	Look for an appropriate clinical question with the aim of addressing an unmet need in cancer imaging. A useful question to ask is, "What is the likely added value of a radiomic model above existing clinical models?"
Are there sufficient data?	Aim for 10 samples per feature following feature reduction, although this is dependent on the research hypothesis and study outcome. It may be more difficult to obtain a significant result with smaller sample sizes (ie, <100).
Is there sufficient data quality?	Measure the data attrition rate: 50% or greater is common. Then ask, "Are there sufficient data after attrition?"
Are the data balanced (time to event)?	Assess the proportion of events occurring and/or lost during follow-up.
Are the data balanced (classification)?	Assess the proportion of samples in each class. If data are very unbalanced, a larger sample size may be needed.
How heterogeneous are the data?	Evaluate differences in the tumor type, scanner used, imaging site, and acquisition protocol performed among the samples. Heterogeneity may introduce confounders, which can be later evaluated after data collection by testing for differences between samples in each of the confounding groups (eg, using a *t* test or ANOVA (analysis of variance) for more than two groups).

Contrast-enhanced imaging yields information about tumor enhancement, vascularity, and heterogeneity (Fig 2) that may not be apparent without the use of contrast material but may incur a cost burden and require particular expertise (eg, ability to perform contrast-enhanced US).

Suitable imaging data that meet the study inclusion and exclusion criteria should be

clearly defined. Standardized imaging protocols (ie, those that use the same vendor or scanner settings for all samples) can be used to reduce unwarranted confounders and noise (4), whereas less rigidly standardized protocols can be used to reflect real-world clinical scenarios.

Once a cohort has been identified, images should be anonymized to remove patient identifiable metadata. However, relevant nonidentifiable image data can be retained. Images should be exported as Digital Imaging and Communication in Medicine (DICOM) files by using a lossless-compressed format to avoid losing potentially informative image features. It is worth speaking with the picture archiving and communication system (PACS) team to enlist help.

Data Curation

Nonimaging and clinical data are typically collated in a repository for analysis, and it is advisable to discuss with the institution's or practice's statistician or data scientist the desired format before data collection. Curation steps to identify missing or incomplete data can then be taken, along with correction of typographic errors or inconsistencies, before merging clinical and radiomic data.

Image Preprocessing

Before feature extraction, the raw image data can be enhanced through a variety of preprocessing steps, which are summarized in Table 3. Although these may improve image quality, care should be taken as they can mask or degrade the radiomic signature and may be better mitigated against by optimizing and standardizing the image acquisition.

Unlike at CT, units of MRI SI are arbitrary, and hence normalization of SI is recommended. Although no consensus currently exists, the z-score is a simple method and is computed by subtracting the mean SI of the region of interest (ROI) from the pixel SI and dividing the result by the standard deviation (19). Bias field correction should also be applied to correct for the spatial field inhomogeneities encountered with MRI (17). Thresholding on voxel Hounsfield units can be applied to CT data to exclude voxels that are assumed to contain noninformative tissues. For example, very low values may correspond to air within the lung and high values to bone or calcification.

As some radiomic feature values are dependent on voxel size (20), images should be resampled to a common spatial resolution for all samples (21). Linear interpolation is generally recommended (18,22).

Motion correction can be used to correct for misregistration, blurring, or motion artifacts

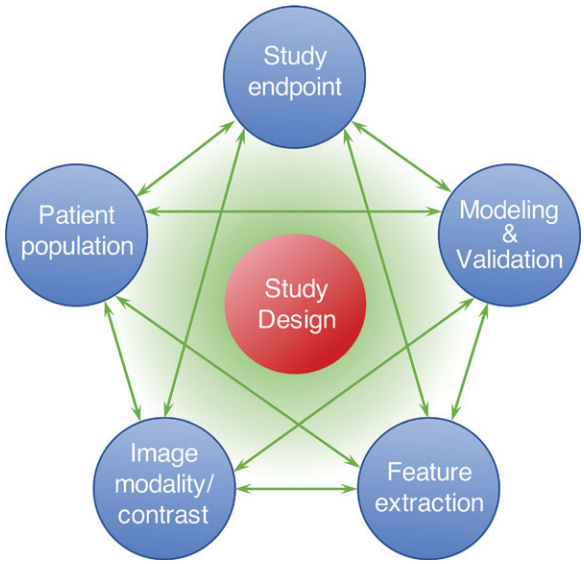


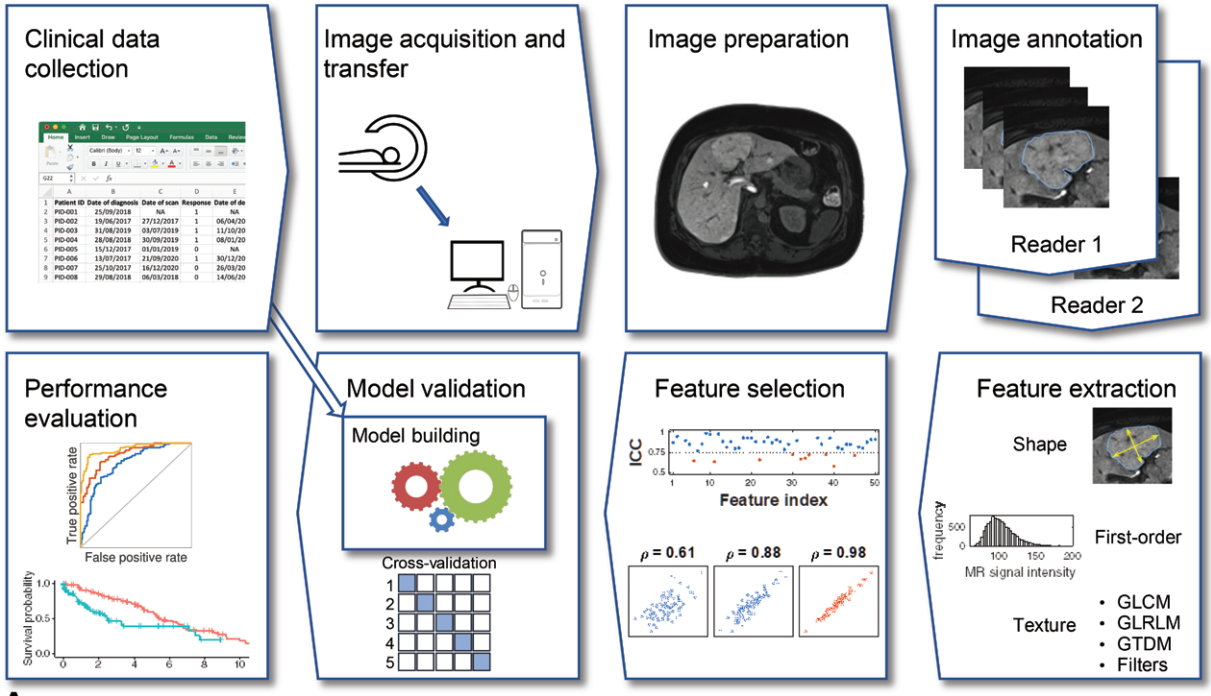
Figure 3. As demonstrated in this diagram, the study design arises by considering the interaction of multiple criteria or activities, including patient population, study endpoint, available imaging and/or clinical data, radiomic feature extraction methodology, and appropriate modeling and validation strategy.

and has been used in four-dimensional CT of lung tumors (23). However, this additional processing has the potential to impact potential radiomic information in the images. The use of motion- control techniques, such as breath holding, is advised as the effect of motion blurring on computed radiomic features is known to be feature dependent (24).

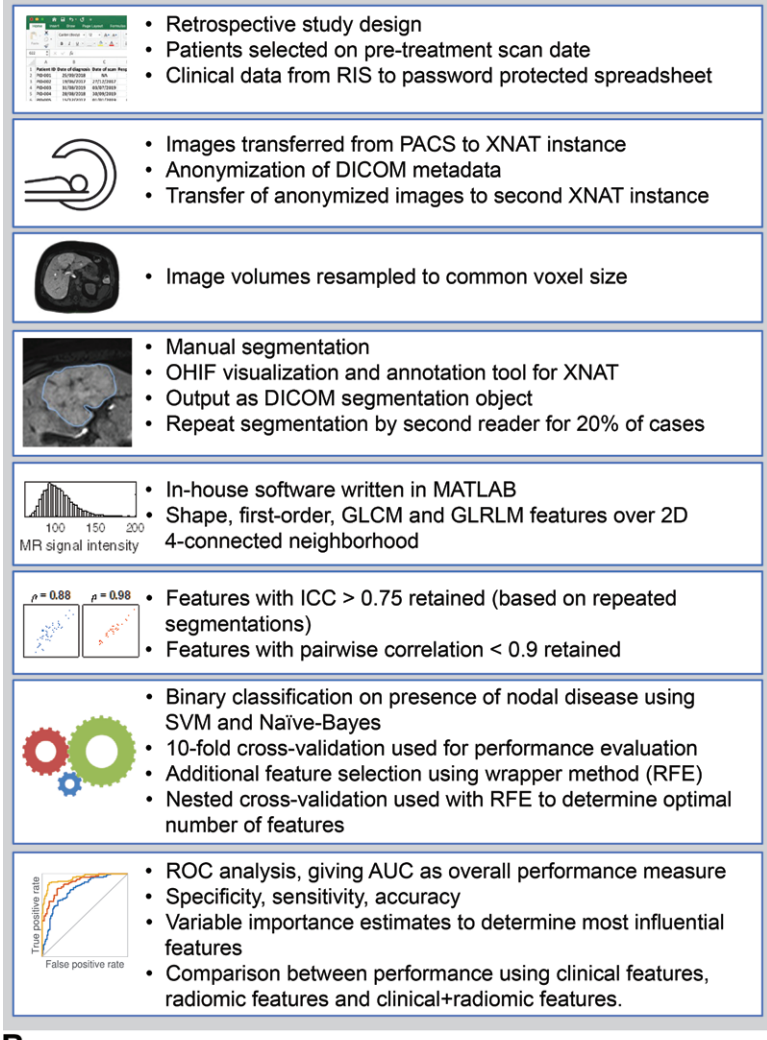
Image filtration can be used before the extraction of features as a preprocessing step to highlight particular image properties. Nonspatial filters increase or decrease the sensitivity of the radiomics features to high- or low-intensity values; examples include taking the square or exponential of the image intensities. Spatial filters increase or decrease the sensitivity of features to particular spatial properties of the image. Examples include Laplacian of Gaussian (LoG) filters, which emphasize areas of rapid change (eg, edge detection) (Fig 5) (25) and wavelet filters, which separate high- and low-spatial-frequency information. The number of radiomics features (and hence datasets) generated with image filtration can become large, so it is typical to try using unfiltered images first.

Segmentation

Segmentation can be performed by drawing ROIs on the tumor, tumor subregions ("habitats"), or peritumoral zones, the choice guided by the research hypothesis. For example, habitat imaging aims to characterize intratumoral spatial heterogeneity by comparison of discreet functional tumor subregions (26), whereas the peritumoral zone



A



B

Figure 4. **(A)** Overview of a typical radiomic workflow that embodies the study design and details the steps involved in taking clinical and imaging inputs all the way through to the study endpoint. **(B)** Details of each stage should be clearly reported to allow meaningful interpretation, discussion, and critique of the study findings. The workflow used in Doran et al (15) is illustrated. The authors investigated the utility of radiomics from multivendor multi-parametric MRI in prediction of lymph node status in patients with breast cancer. *AUC* = area under the curve, *DICOM* = Digital Imaging and Communication in Medicine, *GLCM* = gray-level co-occurrence matrix, *GLRLM* = gray-level run-length matrix, *GTDM* = gray-tone difference matrix, *ICC* = intra-class correlation coefficient, *OHIF* = Open Health Imaging Foundation, *PACS* = picture archiving and communication system, *RIS* = radiology information system, *ROC* = receiver operator characteristic, *RFE* = recursive feature elimination, *SVM* = support vector machine, *2D* = two dimensional, *XNAT* = eXtensible Neuroimaging Archive Toolkit.

Table 3: Image Preprocessing Steps That Can Be Used before Feature Extraction

Preprocessing Step	Note
SI normalization	Recommended for MRI as SI values are arbitrary, although no consensus for this method exists currently
De-noising	Recommended for MRI, which contains Gaussian and Rician noise (16)
Bias field correction	Recommended for MRI to correct unwarranted spatial signal variation due to inherent field inhomogeneities (17)
Image interpolation and resampling	May be necessary to produce imaging data of uniform spatial resolution; interpolation method may impact feature values (18)
Motion correction	May be useful for PET data or dynamic contrast-enhanced (DCE) MRI parametric map data to correct for misregistration or motion artifacts; has potential to impact feature values
Image thresholding	Can be used with CT data to exclude pixels with Hounsfield units outside of a specified range; this may reduce effect of noise, while highlighting desired attenuation values

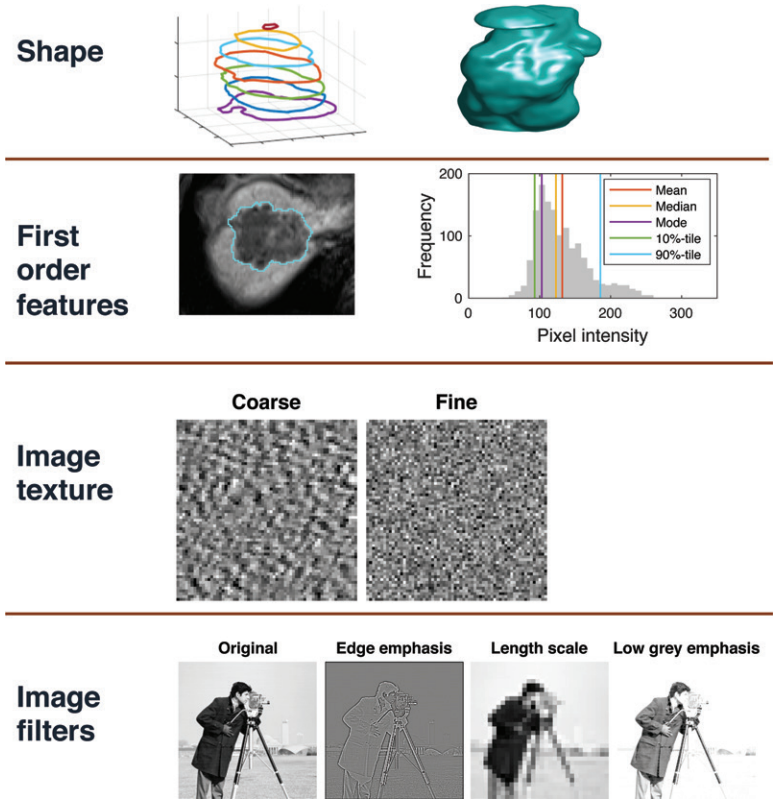


Figure 5. Pictorial overview of the feature classes used in most radiomic studies. Shape or morphologic features can be computed in 2D or 3D views, with 3D analysis being the recommended approach for most studies. First-order features are computed from the distribution of SIs within the ROI and include features such as the mean, median, and mode, which describe the central tendency of the data, and other features such as percentiles, skewness, kurtosis, and entropy, which describe the symmetry and heterogeneity of the distribution. Texture or second-order features consider the joint statistics of two or more voxels, so that in the coarse texture example, neighboring pairs of pixels are likely to have similar gray levels, whereas in the fine texture example, neighboring pixel values are independent. In radiologic images, the statistical dependencies between neighbors can be more complex than in these simple examples, and so features derived from the GLCM, gray-level run-length matrix (GLRLM), and other metrics can be effective for quantifying image texture. Filtering the images to emphasize edges, different length scales, or different gray levels can be used before computing texture features with the aim of sensitizing the features to a wider range of biologic correlates.

may contain information about tumor invasion or host immune response (27). Radiation therapy tumor volume data used for treatment planning can also be used, although these may differ from ROIs specifically drawn for a radiomic analysis.

ROIs can be delineated manually, automatically, or semiautomatically in either two dimensions (2D) (single section) or three dimensions (3D) (multiple sections) (Fig 5). The choice will be determined by available resources and tumor type. Three-dimensional ROIs will capture additional information but can be time consuming to draw when manual delineation is used.

Automatic segmentation is potentially faster and more reproducible (28) and may be required for larger datasets for which manual segmentation is not feasible. However, segmentations should be checked by a radiologist to ensure accuracy. Features can be compared against those obtained after manual segmentation by using the Dice score.

When manual segmentation is used, feature stability should be assessed by performing multiple segmentations of the same tumor with either the same or a different reader performing the delineation.

Table 4: Summary of Radiomic Feature Classes

Feature Class	Examples	Note
Morphologic	Diameter, area, sphericity	Semantic features may represent descriptive scores (eg, small, medium, large). However, there are corresponding morphologic radiomic features that are purely quantitative.
Intensity	Minimum, maximum, mean 10th and 90th percentiles, skewness, kurtosis	First-order features describe properties of the distribution of SIs within an ROI (eg, the minimum, maximum, mean, median, range, standard deviation, and 10th and 90th percentiles of the intensities). Skewness refers to asymmetry of the distribution of values about the mean and can be positive or negative. Kurtosis refers to the tail behavior of the SI distribution, with higher values implying a higher proportion of SI values concentrated toward the tails and a lower proportion toward the mean.
Texture features	Contrast, correlation, entropy, run emphasis, gray-level nonuniformity	Second-order features describe spatial complexity and relationships of SIs between neighboring pixels; often computed from the co-occurrence matrix (GLCM) described by Haralick (8) or the run-length matrix (GLRLM) described by Galloway (29). Other classes include those derived from the gray-level size-zone matrix (GLSZM) (30), gray-level distance-zone matrix (GLDZM) (30), neighborhood gray-tone difference matrix (NGTDM) (31), and neighborhood gray-level dependence matrix (NGLDM) (32).

Feature Extraction

Feature extraction is the final step before model building and validation and involves computing radiomic features from each ROI that will be used in the model. Radiomic features are "handcrafted" in that the algorithms used to generate them are designed or chosen by the data scientist rather than being learned directly from the images, as is found with deep learning approaches. Consequently, it may be possible to interpret the radiomic signature obtained with handcrafted features, whereas deep learned features can suffer from limited explainability.

A wide variety of feature classes exist and are summarized in Table 4. The set of quantitative imaging features is large and is being continually updated and refined. Efforts have been made in standardization such as with the Image Biomarker Standardization Initiative (IBSI) (21), and we recommend that readers refer to this resource for an up-to-date description of features and their properties.

Morphologic features describe geometric properties of the lesion such as volume, diameter, surface area, and elongation. Intensity-based features, also known as first-order features, describe properties of the distribution of intensities within an ROI, where the spatial location of each voxel is ignored. First-order features can be broadly grouped into those that measure the location of the distribution (mean, median, mode, etc), those that measure the spread of the distribution (variance, interquartile range, etc), those that measure the shape of the distribution (skewness, kurtosis, etc), and other features linked to less specific properties of the voxel intensity heterogeneity (entropy, energy, etc). Imaging modalities such as MRI and US typically generate images with arbitrary intensity scaling, and if this is not consistent for all subjects it will be necessary to apply image standardization before calculating first-order features. Features such as skewness are unaffected by image standardization, as they are dependent on the shape of the distribution of intensities rather than their absolute values.

Second-order features, also known as texture features, go beyond first-order features so that the spatial locations as well as the SIs of two or more pixels are used when computing the features. For example, gray-level co-occurrence matrix (GLCM) features consider the SIs of pairs of pixels separated by a given distance and direction, while gray-level size-zone matrix (GLSZM) features consider the sizes of contiguous regions that share the same SI after discretization.

Intensity discretization involves assigning pixels within a given intensity range to a single value or "bin" and is used before calculation of second-order features. Either the bin width or the total number of bins can be specified. Reducing the number of bins (or increasing the bin width) will lead to a loss of image detail but will remove noise (Fig 6). Conversely, increasing the number of bins (or decreasing the bin width) will retain more image detail but will also preserve image noise. Using a fixed bin size maintains the relationship of the "binned" data to the original intensity scale and can be used when the intensity scale is quantitative

(such as CT and PET data). When image intensity units are arbitrary (such as with MRI data), fixing the number of bins (rather than the bin size) is recommended (21). Whichever method is used, it should be the same for all patients.

In addition to the agnostic or quantitative feature classes described, semantic features such as "spiculated" or "enhancing" can also be used as input features to a radiomics model and will be determined by visual inspection. These features will typically be categorical (eg, small, large, hyperenhancing) rather than numerical.

Model Building

Once clinical and radiomic data are collected and curated, statistical models are fitted to predict study endpoints, such as tumor type or survival time. A typical model uses input features (including the radiomic features described previously and clinical features such as tumor markers or lymph node status) in addition to target data that the model aims to predict, such as benign versus malignant or risk of recurrence. The final performance and generalizability of models discovered from a radiomic analysis is determined by validating the model on new test data (33,34).

The hold-out method uses a training set to develop the model, and a validation set to estimate future performance on new data. To avoid biasing the model performance, the validation data should be shielded from the model training process, and the final validation only performed once. Ideally, validation data should be obtained from another institution, but this is not always possible. Splitting single-institution data into training and validation sets is often more practical and can be done randomly, temporally (by using the most recent cases as validation data), or by choosing a similar class proportion (eg, benign versus malignant) in the training and validation datasets, known as stratified sampling.

Once training and validation datasets have been established, it is important to verify that the feature distributions between the two datasets are similar. This is to ensure that any informative patterns obtained in the training data will also be present in the validation data. Independent univariate testing of each feature is typically performed, and useful tests include the Mann-Whitney *U* test (equality of the medians in the two datasets), and the Komogorov-Smirnov or Shapiro-Wilk test (equality of the distributions of the two datasets). These tests do not use the outcome data (they are referred to as unsupervised) and therefore do not violate the rule that the validation data should only be used for model testing.

While hold-out validation is the most straightforward approach, it works less well with small datasets (<100–200 samples) because uncertainty of the performance in the validation dataset will be large, and the diversity of the training data may be insufficient to discover a robust model. If obtaining more data is unfeasible, and in the case of smaller studies, cross-validation can be used to estimate performance.

With K-fold cross-validation, data are partitioned into K folds (typically 3–10), then K-1 folds are used to train the model, and the remaining fold is reserved to test the model. In this way, K separate models are trained, in which each fold plays the role of the test set. The final performance estimate is the average over all the folds, and the standard error of the performance can be estimated by using the standard deviation over the folds. This is useful when comparing different models and reflects the robustness of the model.

Many models have tuning parameters, and optimizing these parameters can be crucial for good performance. Unlike the model parameters, tuning parameters cannot be learned directly from training data. Poorly tuned parameters can lead to over- or underfitting of the training data—overfitting leads to poor performance in the validation data compared with the training data, and underfitting occurs when the model is unable to capture important features in the training data (Fig 7). Split-validation (equivalent to hold-out validation) and cross-validation can be used for optimizing the tuning parameters, and this enables tuning parameters to be found that balance between over- and underfitting. Similar validation approaches can be used to select between candidate models.

Feature Stability

When manual segmentation is performed, it is important to reject radiomic features that are particularly sensitive to interreader variations in the ROI. This is evaluated by repeating tumor segmentations for a subset of patients by one or more readers. The intraclass correlation coefficient (ICC) can be used to reject nonreproducible features (35,36) below a threshold ICC.

While patients used for measuring reproducibility can be selected from the whole dataset, when hold-out testing is used it is convenient to select patients who are in the training data. In this case, the ICC threshold for feature rejection can be treated as a model parameter and optimized as a tuning parameter, but when this is not performed, ICC thresholds in the range of 0.75–0.9 are typical. Feature stability is also influenced by fluctuations in patient factors, including positioning; if possible, test-retest images on a subset of patients should be obtained. This is often feasible with MRI studies but can be more difficult for images obtained involving

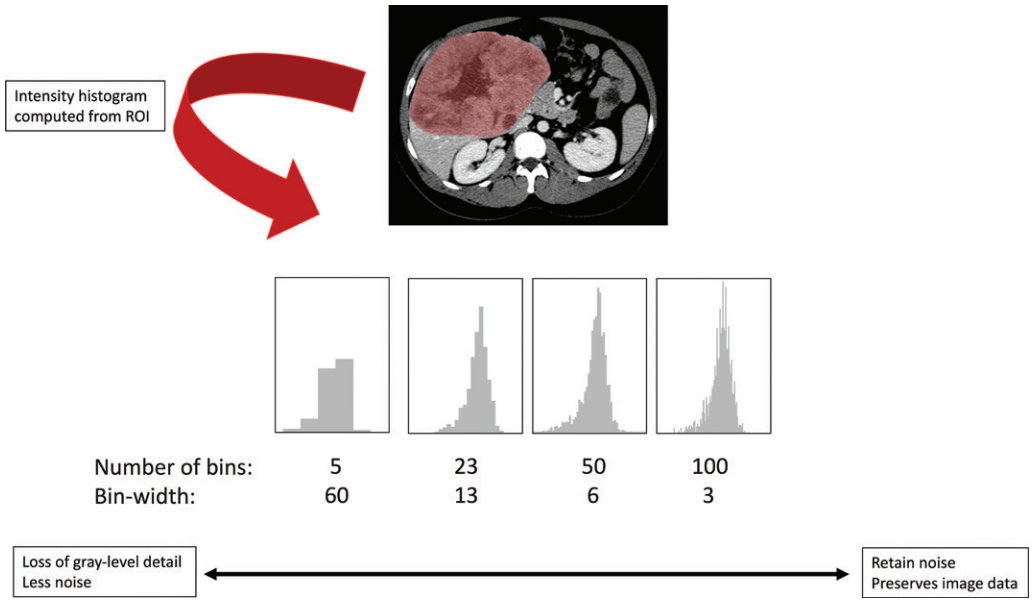


Figure 6. SI discretization involves assigning pixels within a given SI range to a single value or bin and is used before calculation of second-order features. In this diagram, the SI histogram is derived from an ROI encompassing a hepatic tumor with varying bin size (or bin width). Increasing the bin size or decreasing the number of bins may cause loss of image detail but reduces noise, whereas decreasing the bin size or increasing the number of bins preserves image detail at the expense of image noise. The choice of image modality and SI range will define the method of discretization.

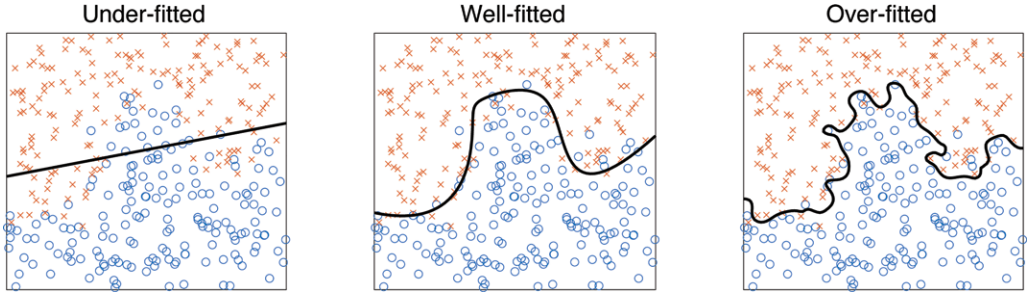


Figure 7. Example 2D classification tasks show the impact of under- and overfitting. In the case of underfitting, the linear model fits a straight line and does not have the capacity to capture the nonlinear (curved) nature of the decision boundary, and so its classification performance on both the training and the test data will be suboptimal. In the case of overfitting, the model is insufficiently constrained and tends to generate a complex decision boundary that is overly influenced by noise. In this case, the performance in the training data will be good but will worsen when evaluated on independent test data. Many machine learning models have tuning parameters that can be adjusted to give models at both ends of this spectrum, and so optimizing the tuning parameters (typically using cross-validation techniques) is necessary to produce a well-fitted model.

ionizing radiation and is usually not possible in retrospective studies.

Univariate Feature Discovery

In radiomic studies, it is uncommon for a single feature to perform well enough to be used on its own, but univariate models (those that contain only one feature) are nevertheless useful as a benchmark baseline performance for comparison with more complex multivariate models (those that contain multiple features). For binary classification tasks (in which data are categorized into two groups [eg, benign vs malignant]) the area under the receiver operating characteristic curve is a suitable metric to rank the classification performance of each feature when used alone, and the Mann-Whitney *U* test can be used to test whether the model performs better than chance alone. As classification performance will be measured for each feature, multiple comparisons correction of the *P* values should be performed by using Bonferroni correction or false–discovery rate methods such as Benjamini-Hochberg and Benjamini-Yekutieli corrections (37,38).

Feature Selection and Dimensionality Reduction

Multivariate models often perform better when feature selection or dimensionality reduction is applied because this tends to remove noise and

reduces redundancy (the number of features that do not add any additional information to the model). A range of approaches are outlined in Table 5 and described in more detail in this section. A key consideration when choosing a feature selection technique is the impact on interpretability of the final model.

Correlated features can be reduced by using pairwise correlation statistics (such as Pearson correlation) to remove features that are correlated above a threshold (eg, 0.8). This is performed without knowledge of the outcome data and is done iteratively starting with the pair with the highest correlation. For each pair, the feature with the highest average correlation with the remaining features is rejected. To improve interpretability in addition to stability, we have developed an extension to this technique in which the classes of highly correlated features (ie, shape, first order, and texture) are used to determine which feature should be removed (15). For example, if a first-order and a texture feature are correlated, then the first-order feature is retained, and if a shape and a first-order feature are correlated, then the shape feature is retained. This results in a set of features with reduced redundancy and tends toward simpler interpretations.

Dimensionality reduction techniques aim to retain the informative components of features with a smaller overall number of variables. For example, the majority of "useful" information contained in 100 features is represented in one or two new variables that comprise combinations of the features. In this case, the dimensions have been reduced from 100 to one or two. Widely used examples include principle component analysis (PCA), independent component analysis, kernel PCA (39), and autoencoders (40). A key limitation is that variables obtained following feature reduction suffer from limited explainability since they are influenced by a combination of many or all of the input features.

Feature selection methods make use of the target data, and these can be divided into three types: filter, wrapper, and embedded methods.

Filter methods use statistics derived from each input feature and the target data to rank and select the input features and are applied to the training data before the model fitting. They are supervised (as they make use of the target data), and care should be taken to avoid data leakage from the validation data. Possible statistics include the *t*-statistic, Mann-Whitney *U* test, Fisher score, joint mutual information, maximum relevancy minimum redundancy, and mutual information (41).

Wrapper methods combine the chosen multivariate model with a feature ranking function that is used iteratively to remove low-ranking features. To avoid overfitting, the ranking should be computed by using cross-validation or split validation on the training data. Recursive feature elimination is a popular wrapper method and is available in most statistical packages.

Embedded methods take an existing statistical model (eg, logistic regression) and add a term (known as the regularization term) that has the effect of shrinking model parameters that are associated with noninformative features to values at or near zero. This simplifying property is advantageous when attempting to interpret the final model. Examples include Least Absolute Shrinkage and Selection Operator (LASSO) (42), ridge, and elastic net regularization (43). Embedded methods have one or more tuning parameters, and these should be optimized in the training data by using cross-validation or split-validation.

Multivariate Models

Multivariate models refer to those that use multiple input variables and are frequently used in radiomic studies. The workhorse models for radiomics studies are classification and time-to-event (survival) models (Table 6).

Classification models generate boundaries between the data to separate them into discrete groups (Fig 7). These are referred to as decision boundaries, and data are classified on the basis of which side of the boundary they are located. A widely used group of classification models generate linear boundaries (ie, a straight line) or quadratic boundaries (a curve). These include linear discriminant analysis (LDA) and Gaussian naïve Bayes and quadratic discriminant analysis. Logistic regression is a related technique that (like LDA) generates a linear decision boundary, but unlike LDA, data points that are far from the boundary have a reduced effect on the location of the boundary. These classification models have the advantage that they do not have any tuning parameters but the disadvantage that they can only generate linear (or quadratic) decision boundaries, which may result in underfitting if the true boundary separating classes is not simply a straight line or quadratic function. These techniques can be used in combination with all three feature selection methods described previously. Logistic regression with LASSO regularization is a widely used example of this and has the advantage that the model parameters can be interpreted as odds-ratios, and the regularization tends to remove noninformative features, which aids model interpretation.

When the data require a more complex decision boundary, nonlinear classifiers such as support vector machines, relevance vector machines, random forests, and neural network classifiers

Table 5: Methods for Feature Selection and Reduction

Technique	Description	Advantages	Disadvantages
Manually remove features	Selecting a feature or features to remove on the basis of priori considerations (eg, if a study involves very small tumors, texture features are unlikely to be informative)	Simple to apply Does not use the target features	Introduces selection bias that is not validated
Feature correlation	Removes features on the basis of their individual correlation with each other (known as pair-wise correlation) above a chosen threshold (eg, 0.8); usually performed iteratively	Simple to apply Does not use target features	Correlation threshold needs careful selection and will influence performance
Feature stability	Assess temporal stability in the test-retest setting or segmentation stability following repeat segmentations	More likely that the performance estimate will remain true when the model is deployed in a real-world scenario	Unstable features could be informative if efforts were made to improve stability
Dimensionality reduction	Examples include principle component analysis (PCA) and independent component analysis (which use linear transformations of the input features), kernel PCA, and autoencoders (which use nonlinear transformations); informative components of features are retained with fewer overall variables	Interactions (eg, correlations) between features are automatically accounted for; high capacity to reject noise and retain informative signal	Lack of interpretability of features after dimensionality reduction
Filter methods	Select relevant features on the basis of statistical testing by using outcome variable; top-performing features are retained	Applied as a preprocessing step before model building	Parameter selection is not influenced by the model used, so it may reject features that would be informative for a given model
Wrapper methods	Recursively select or reject features on the basis of model performance	Combines selection and model fitting so that features that are relevant for a particular model are retained; features are not modified, so interpretation of the final model is feasible	The number of retained features should be determined by using cross-validation, leading to slower and more complex algorithms
Embedded methods	Augment the model with a regularization term that leads to sparse models (ie, that have coefficients that are close to zero)	Combines selection and model fitting so that features that are relevant for a particular model are retained; features are not modified, so interpretation of final model is feasible	Regularization includes a tuning parameter that should be determined by using cross-validation, leading to slower more complex algorithms

may be appropriate (33,44). These algorithms can generate more complex boundaries between classes (compared with those of a linear or quadratic function) and have tuning parameters that can have a dramatic effect on performance, and so cross-validation or split-validation on the training data should be used for tuning parameter optimization.

Evaluating the performance of classification models learned from data is a crucial aspect of the development of a radiomic signature. Some of the more widely used metrics and their uses are outlined in Appendix E2.

Time-to-event models widely used in radiomics studies include Cox regression and random forest survival models. Both models account for data censoring. In radiomic studies with a large number of input features, Cox regression with LASSO regularization can be effective at generating a risk signature with a small number of nonzero features (45). Performance assessment of time-to-event

Table 6: Advantages and Disadvantages of Widely Used Classification and Time-to-Event Models

Model	Advantages	Disadvantages
Classification		
Linear/quadratic discriminant analysis	Simple model directly estimated from the data, no tuning parameters; probabilistic output available	Decision boundary constrained to be a straight line (or quadratic curve); suboptimal for non-Gaussian class distributions and data with extreme values
Gaussian Naïve Bayes	Simple model directly estimated from the data, no tuning parameters; probabilistic output available	Assumes features are independent
Logistic regression	Directly gives probabilistic output; less affected by extreme values not near the decision boundary; regularization can be used for embedded feature selection, which aids interpretation	Decision boundary constrained to be a straight line; can overfit when the number of features (input dimensionality) is high
Support (and relevance) vector machines	Can learn nonlinear decision boundaries; work well in problems with high-dimensional input data	More complex algorithms with slower run-times; involve tuning parameters that almost always require optimizing to give good performance; interpreting the final model can be difficult
Random forest classifier	Can learn nonlinear decision boundaries; naturally robust, as each tree is trained by using a subset of the data	May require more data than support (and relevance) vector machines to learn complex decision boundaries; involves tuning parameters that should be optimized for good performance; results can be difficult to interpret if the tree depths are >3
Time-to-event		
Cox regression	Widely used and understood; ease of interpretation, as the model gives hazard ratios for each input parameter; regularization can be used for embedded feature selection; accounts for data censoring	Assumes proportional hazards and linear relationship between input features and hazard
Random forest survival	Can model nonlinear and nonproportional survival effects; accounts for data censoring	Contains tuning parameters that require optimization; may require more data than the Cox regression

models broadly falls into two types: prediction accuracy at a given time point or accuracy at predicting risk for the whole survival curve (46). Common metrics for assessing these are outlined in Appendix E2.

Software

The main initial consideration when choosing radiomic software is whether to use commercial or noncommercial software. Noncommercial applications tend to be free, rapidly evolving, and reflective of the latest research trends. Commercial applications are not free but may be more stable, come with technical support, and are potentially a "black box." As with all scientific software, users should consider the maturity level of the chosen package, documentation available, previous use in the literature, and potential for support from the individual or organization developing it. Additional radiomic-specific considerations include picture archiving and communication system (PACS) integration, segmentation tools, radiomic features supported, preprocessing, and model building. If local expertise is available, consider implementing an in-house pipeline that may be optimized to local systems. A large number of noncommercial software applications (Appendix E3) have been developed, and many are freely available for public download.

At present, there is limited choice of commercial software in this rapidly developing field. This likely represents a combination of the low potential for revenue where many open-source solutions already exist and the high barrier for developing software as a medical device for sale to the health care market. It is important to note that reproducibility is not guaranteed simply by using IBSI-tested software but also relies on harmonizing certain settings (which do not necessarily correspond to the defaults of the software) and maintaining consistency in versions of each software platform (47).

Manuscript Writing

The interpretation of findings from radiomic studies requires detailed knowledge of the vari-

ous steps performed during the study design, and it is crucial that these are clearly outlined when preparing a manuscript. To aid authors and to provide a framework for manuscript writing, there are various radiomic- and artificial intelligence–specific checklists, reporting guides, and radiomic quality scores that can be referred to (4,21,48,49), in addition to artificial intelligence extensions of familiar guidelines such as TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) (50), CONSORT (Consolidated Standards of Reporting Trials) (51), and SPIRIT (The Standard Protocol Items: Recommendations for Interventional Trials) (52). These can help assist with manuscript preparation, with insights into how manuscripts will be assessed at peer review. To address the challenge of standardization in radiomics, it is important to observe recognized nomenclature, for example, that collated by the IBSI (21).

Processing and acquisition parameters should be specified for all stages of the study, in addition to software details and version numbers. It has been proposed that liberal use of supplementary materials to include imaging protocols, examined images, segmentations, formulas for feature extraction, and code of radiomic models is encouraged (4). Where it is not possible to present patient-specific data, computed values from a digital phantom (53) can be used and compared with validated tolerance levels (54).

Future Directions and Challenges

Although there has been an exponential increase in the number of radiomic publications, routine clinical implementation is yet to occur (55,56).

Key obstacles include noncompliance with machine learning best practices, standardization of the radiomic workflow, and clear reporting of study methodology. Only then can models be validated, preferably prospectively on external real-world data, including multivendor images and a variety of acquisition protocols.

Data curation and quality and adequate sample sizes are crucial in meeting these challenges. However, curation of large datasets is resource intensive, and accrual of sufficient data from multiple institutions can be challenging. Data sharing can help address these challenges. However, hurdles remain (57), including but not limited to ethical and legal considerations, data value, intellectual property, and resource availability.

Finally, although radiomics is largely a data-driven exercise, a deeper understanding of the biologic meaning of any derived radiomic signatures is required before results gain wider acceptance (6).

Conclusion

Radiomic applications in oncology include diagnosis, prognostication, and prediction of clinical outcomes. It is a multidisciplinary field, encompassing radiologists and data and imaging scientists. A variety of challenges exist, including a need for standardization across all stages of the workflow and prospective validation across multiple sites using real-world heterogeneous datasets. This article provides multiple learning points to improve study design and execution and to enhance translation of radiomics into clinical practice.

Acknowledgment.—The authors would like to thank Naami Mcaddy, MBBS, for their assistance with reviewing the manuscript.

Disclosures of Conflicts of Interest.—**S.J.D.** *Activities related to the present article:* post funded via a grant from Cancer Research UK. *Activities not related to the present article:* post funded via a grant from Cancer Research UK. *Other activities:* disclosed no relevant relationships. **N.P.** *Activities related to the present article:* disclosed no relevant relationships. *Activities not related to the present article:* stock/stock options in MRIcons. *Other activities:* disclosed no relevant relationships. **D.M.K.** *Activities related to the present article:* disclosed no relevant relationships. *Activities not related to the present article:* institutional support from NIHR Challenge Award and payment for lectures from Bayer Healthcare. *Other activities:* disclosed no relevant relationships.

References

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. Radiology 2016;278(2):563–577.
2. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. Eur J Cancer 2012;48(4):441–446.
3. Kumar V, Gu Y, Basu S, et al. Radiomics: the process and the challenges. Magn Reson Imaging 2012;30(9):1234–1248.
4. 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.
5. Aerts HJWL, Velazquez ER, Leijenaar RTH, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun 2014;5(1):4006.
6. Tomaszewski MR, Gillies RJ. The Biological Meaning of Radiomic Features. Radiology 2021;298(3):505–516 [Published correction appears in Radiology 2021;299(2):E256.]
7. Larue RTHM, Defraene G, De Ruysscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. Br J Radiol 2017;90(1070):20160665.
8. Haralick RM, Shanmugam K, Dinstein I. Textural Features for Image Classification. IEEE Trans Syst Man Cybern 1973;SMC-3(6):610–621.
9. O'Connor JPB. Rethinking the role of clinical imaging. Elife 2017;6:e30563.
10. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 2018;15(2):81–94.
11. Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis: I—Background, goals, and general strategy. J Clin Epidemiol 1995;48(12):1495–1501.
12. Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis: II—Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48(12):1503–1510.
13. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per

variable in logistic regression analysis. J Clin Epidemiol 1996;49(12):1373–1379.
14. Julious SA. Sample size of 12 per group rule of thumb for a pilot study. Pharm Stat 2005;4(4):287–291.
15. Doran SJ, Kumar S, Orton M, et al. "Real-world" radiomics from multi-vendor MRI: an original retrospective study on the prediction of nodal status and disease survival in breast cancer, as an exemplar to promote discussion of the wider issues. Cancer Imaging 2021;21(1):37.
16. Gudbjartsson H, Patz S. The Rician distribution of noisy MRI data. Magn Reson Med 1995;34(6):910–914.
17. Vovk U, Pernuš F, Likar B. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans Med Imaging 2007;26(3):405–421.
18. Larue RTHM, van Timmeren JE, de Jong EEC, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. Acta Oncol 2017;56(11):1544–1553.
19. Ellingson BM, Zaw T, Cloughesy TF, et al. Comparison between intensity normalization techniques for dynamic susceptibility contrast (DSC)-MRI estimates of cerebral blood volume (CBV) in human gliomas. J Magn Reson Imaging 2012;35(6):1472–1477.
20. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys 2017;44(3):1050–1062.
21. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. http://arxiv.org/abs/1612.07003. Published 2016. Accessed June 27, 2021.
22. Lehmann TM, Gönner C, Spitzer K. Survey: interpolation methods in medical image processing. IEEE Trans Med Imaging 1999;18(11):1049–1075.
23. Du Q, Baine M, Bavitz K, et al. Radiomic feature stability across 4D respiratory phases and its impact on lung tumor prognosis prediction. PLoS One 2019;14(5):e0216480.
24. Yip S, McCall K, Aristophanous M, Chen AB, Aerts HJWL, Berbeco R. Comparison of texture features derived from static and respiratory-gated PET images in non-small cell lung cancer. PLoS One 2014;9(12):e115510.
25. Ganeshan B, Miles KA. Quantifying tumour heterogeneity with CT. Cancer Imaging 2013;13(1):140–149.
26. Napel S, Mu W, Jardim-Perassi BV, Aerts HJWL, Gillies RJ. Quantitative imaging of cancer in the postgenomic era: Radio(geno)mics, deep learning, and habitats. Cancer 2018;124(24):4633–4649.
27. Colleoni M, Rotmensz N, Maisonneuve P, et al. Prognostic role of the extent of peritumoral vascular invasion in operable breast cancer. Ann Oncol 2007;18(10):1632–1640.
28. Parmar C, Rios Velazquez E, Leijenaar R, et al. Robust Radiomics feature quantification using semiautomatic volumetric segmentation. PLoS One 2014;9(7):e102107.
29. Galloway MM. Texture analysis using gray level run lengths. Comput Graph Image Process 1975;4(2):172–179.
30. Thibault G, Angulo J, Meyer F. Advanced statistical matrices for texture characterization: application to cell classification. IEEE Trans Biomed Eng 2014;61(3):630–637.
31. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans Syst Man Cybern 1989;19(5):1264–1274.
32. Sun C, Wee WG. Neighboring gray level dependence matrix for texture classification. Comput Vis Graph Image Process 1983;23(3):341–352.
33. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. New York, NY: Springer, 2001.
34. Cawley GC, Talbot NLC. On over-fitting in model selection and subsequent selection bias in performance evaluation. J Mach Learn Res 2010;11:2079–2107.
35. Zwanenburg A, Leger S, Agolli L, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep 2019;9(1):614.
36. McHugh DJ, Porta N, Little RA, et al. Image Contrast, Image Pre-Processing, and T_1 Mapping Affect MRI Radiomic Feature Repeatability in Patients with Colorectal Cancer Liver Metastases. Cancers (Basel) 2021;13(2):240.
37. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol 1995;57(1):289–300.
38. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat 2001;29(4):1165–1188.
39. Schölkopf B, Smola A, Müller KR. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural Comput 1998;10(5):1299–1319.
40. Hinton G, Zemel R. Autoencoders, minimum description length, and Helmholtz free energy. In: Proc 6th Int Conf Neural Inf Process Syst. San Francisco, CA: Morgan Kaufmann Publishers, 1993; 3–10.
41. Papanikolaou N, Matos C, Koh DM. How to develop a meaningful radiomic signature for clinical use in oncologic patients. Cancer Imaging 2020;20(1):33.
42. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. J R Stat Soc Ser B Methodol 1996;58(1):267–288.
43. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Methodol 2005;67(2):301–320.
44. Tipping M. Sparse Bayesian Learning and the Relevance Vector Machine. J Mach Learn Res 2001;1:211–244.
45. Tibshirani R. The lasso method for variable selection in the Cox model. Stat Med 1997;16(4):385–395.
46. Rahman MS, Ambler G, Choodari-Oskooei B, Omar RZ. Review and evaluation of performance measures for survival prediction models in external validation settings. BMC Med Res Methodol 2017;17(1):60.
47. 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.
48. Mongan J, Moy L, Kahn CE Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. Radiol Artif Intell 2020;2(2):e200029.
49. Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: A brief guide for authors, reviewers, and readers-from the Radiology Editorial Board. Radiology 2020;294(3):487–489.
50. Park JE, Kim D, Kim HS, et al. Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 2020;30(1):523–536.
51. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 2020;26(9):1364–1374.
52. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ 2020;370:m3210.
53. Lambin P. Radiomics Digital Phantom. CancerData 2016. https://doi.org/10.17195/candat.2016.08.1. Published 2016. Accessed June 27, 2021.
54. 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.
55. Pinto Dos Santos D, Dietzel M, Baessler B. A decade of radiomics research: are images really data or just patterns in the noise?. Eur Radiol 2021;31(1):1–4.
56. Langlotz CP, Allen B, Erickson BJ, et al. A Roadmap for Foundational Research on Artificial Intelligence in Medical Imaging: From the 2018 NIH/RSNA/ACR/The Academy Workshop. Radiology 2019;291(3):781–791.
57. Budin-Ljøsne I, Burton P, Isaeva J, et al. DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public Health Genomics 2015;18(2):87–96.