European Journal of Nuclear Medicine and Molecular Imaging (2019) 46:2638–2655
https://doi.org/10.1007/s00259-019-04391-8

ORIGINAL ARTICLE

Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis

Alex Zwanenburg [1,2,3,4]

Received: 31 May 2019 / Accepted: 4 June 2019 / Published online: 25 June 2019

Abstract

Radiomics in nuclear medicine is rapidly expanding. Reproducibility of radiomics studies in multicentre settings is an important criterion for clinical translation. We therefore performed a meta-analysis to investigate reproducibility of radiomics biomarkers in PET imaging and to obtain quantitative information regarding their sensitivity to variations in various imaging and radiomics-related factors as well as their inherent sensitivity. Additionally, we identify and describe data analysis pitfalls that affect the reproducibility and generalizability of radiomics studies. After a systematic literature search, 42 studies were included in the qualitative synthesis, and data from 21 were used for the quantitative meta-analysis. Data concerning measurement agreement and reliability were collected for 21 of 38 different factors associated with image acquisition, reconstruction, segmentation and radiomics-specific processing steps. Variations in voxel size, segmentation and several reconstruction parameters strongly affected reproducibility, but the level of evidence remained weak. Based on the meta-analysis, we also assessed inherent sensitivity to variations of 110 PET image biomarkers. SUV_{mean} and SUV_{max} were found to be reliable, whereas image biomarkers based on the neighbourhood grey tone difference matrix and most biomarkers based on the size zone matrix were found to be highly sensitive to variations, and should be used with care in multicentre settings. Lastly, we identify 11 data analysis pitfalls. These pitfalls concern model validation and information leakage during model development, but also relate to reporting and the software used for data analysis. Avoiding such pitfalls is essential for minimizing bias in the results and to enable reproduction and validation of radiomics studies.

Keywords Positron emission tomography · Radiomics · Reproducibility · Meta-analysis · Systematic review · Machine learning

Introduction

Positron-emission tomography (PET) imaging is increasingly used in a quantitative manner. This requires that intensity values, usually presented as standardized uptake values (SUV), can be compared between repeated measurements, between different scanners, as well as between centres in multicentre trials. In other words, PET imaging should be

This article is part of the Topical Collection on Advanced Image Analyses (Radiomics and Artificial Intelligence).

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00259-019-04391-8) contains supplementary material, which is available to authorized users.

✉ Alex Zwanenburg
alexander.zwanenburg@nct-dresden.de

[1] OncoRay – National Center for Radiation Research in Oncology, Faculty of Medicine and University Hospital Carl Gustav Carus, Helmholtz-Zentrum Dresden – Rossendorf, Technische Universität Dresden, Dresden, Germany

[2] National Center for Tumor Diseases (NCT), Partner Site Dresden, Dresden, Germany

[3] German Cancer Research Center (DKFZ), Heidelberg, Germany

[4] German Cancer Consortium (DKTK), Partner Site Dresden, Dresden, Germany

repeatable and reproducible [1]. In the context of radiomics, repeatability and reproducibility are extended beyond the measurement of mean or maximum SUV to a larger set of image biomarkers. Reproducible image biomarkers do not by themselves guarantee that a radiomics study is reproducible [2]. Most radiomics studies involve data analysis to create prediction models. Several pitfalls need to be avoided to ensure that a model can be successfully used and validated.

Thus, this article is divided into two parts. In the first part, reproducibility of image biomarkers is discussed on the basis of a systematic literature review with a meta-analysis. The second part addresses data analysis and its pitfalls.

Reproducibility of image biomarkers

The most commonly studied image biomarkers in nuclear medicine are mean and maximum SUV of the PET tracer ^{18}F-FDG [3]. Factors that affect the reproducibility of these biomarkers are well understood [4]. Consequently, harmonization guidelines have been formulated to ensure good reproducibility of mean and maximum SUV in multicentre studies [5]. All of the factors that affect mean and maximum SUV also affect the wider range of image biomarkers used in radiomics. In addition, factors related to radiomics-specific image processing and image biomarker computation play a role. Figure 1 shows an overview of a typical radiomics workflow and the factors involved at various stages in the process.

Multiple recent literature reviews have offered a qualitative overview of these factors [6–9]. In this article, we aggregated and quantified evidence for the influence of variability in the various factors on image biomarker reproducibility and determined the inherent sensitivity of image biomarkers to variability through a systematic review and meta-analysis of the available literature [10, 11]. The methodology of the systematic review is presented fully in Supplementary note 1. In brief, PubMed was searched for reports involving PET or SPECT imaging that focused on the repeatability and reproducibility of image biomarkers. Additional reports were added based on recent reviews [6–9]. After screening abstracts, 57 full-text articles were assessed for eligibility, and 42 were included in the qualitative synthesis. No reports on SPECT imaging were found. ^{18}F-FDG was used as the PET tracer in most (36) studies. Results from studies performed in human subjects or anthropomorphic phantoms were eligible for meta-analysis when Bland-Altman metrics [12] and/or within-subject coefficients of variation (CV) were used to assess agreement, or when intraclass correlation coefficients (ICC) [13] were used to assess reliability. Agreement and reliability are not the same concept, and were therefore assessed separately [14, 15]. Values for proportional variability (PV) of the Bland-Altman analysis, CV and ICC were either extracted directly from eligible reports, or requested from the investigators when this was not possible. Data from 21 reports were included in the meta-analysis. The PRISMA flow diagram is shown in Supplementary Fig. 1. The factors and the corresponding studies are presented in Table 1.

Meta-analysis

For the meta-analysis, metric values were transformed using a logarithmic transformation before fitting a linear model to estimate coefficient values for each factor and each image biomarker. The 95% credibility interval was likewise determined for each coefficient. After fitting, coefficients were transformed back to the original scale, where they acted as multipliers. As a multiplier, a value of 1.0 is optimal, as it indicates lack of sensitivity due to variation in a factor or lack of inherent sensitivity of an image biomarker. Higher values indicate increasing sensitivity and decreasing reproducibility. Moreover, the quality of the evidence for each coefficient was assessed using a scoring system based on the number of involved studies and the study quality. For this purpose, the quality of each study was assessed using a list of ten items, as shown in Supplementary note 4. The quality of evidence for

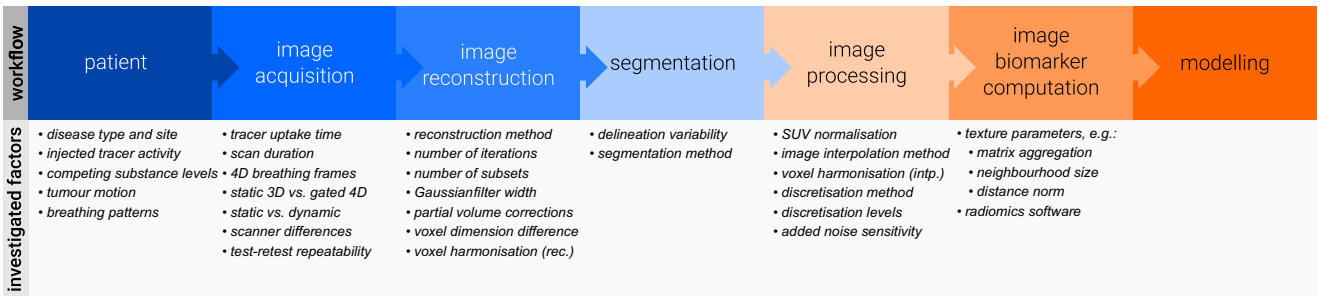

Fig. 1 Typical PET imaging and radiomics workflow. Each step has associated factors which have been investigated for their effect on image biomarker reproducibility. Note that most radiomics studies to date have been retrospective. In such datasets, variation in factors for the steps from patient to reconstruction – that are part of conventional PET imaging workflows – cannot be directly mitigated. Modelling has no effect on reproducibility of image biomarkers, although it does affect reproducibility and generalizability of the study (*rec.* reconstructed, *intp.* interpolated)

Table 1 Overview of factors affecting reproducibility of image biomarkers that were assessed as part of the literature review

Factor	Number of studies selected				References
	Qualitative synthesis		Meta-analysis		
	Total	High quality	Total	High quality	
Patient					
Disease type and site	–	–	–	–	–
Injected tracer activity	1	1	0	0	[16]
Competing substance levels	–	–	–	–	–
Tumour motion	2	1	0	0	[17, 18]
Breathing patterns	1	1	0	0	[17]
Image acquisition					
Tracer uptake time	2	2	0	0	[19, 20]
Scan duration	2	0	2	0	[21[a], 22[a], 23, 24]
4D breathing frames	3	1	1	0	[18[a], 25, 26]
Static 3D vs. gated 4D	4	1	2	1	[18, 22[a], 25[a], 26]
Static vs. dynamic	1	1	1	1	[27]
Scanner differences	1	1	0	0	[28]
Test–retest repeatability	8	2	7	2	[29[a], 30, 31[a], 32[a], 33[a], 34[a], 35[a], 36[a]]
Image reconstruction					
Reconstruction method	10	4	4	2	[21[a], 23, 24, 30, 31[a], 35, 37[a], 38, 39, 40[a]]
Number of iterations	6	3	3	2	[16, 21[a], 24, 31[a], 38, 40[a]]
Number of subsets	2	1	1	1	[24, 31[a]]
Gaussian filter width	8	3	3	2	[16, 21[a], 24, 30, 31[a], 38, 40[a], 41]
Partial volume corrections	1	1	1	1	[42[a]]
Voxel dimension difference	7	3	3	2	[21[a], 24, 30, 31[a], 38, 40[a], 43]
Voxel harmonization (rec.)	1	1	1	1	[25[a]]
Segmentation					
Delineation variability	6	3	3	1	[19, 20, 32[a], 44[a], 45, 46[a]]
Segmentation method	14	9	5	4	[17, 25, 31[a], 35, 37[a], 39, 41, 42[a], 43, 45, 47[a], 48, 49[a], 50]
Image processing					
SUV normalization	–	–	–	–	–
Image interpolation method	1	0	0	0	[51]
Voxel harmonization (intp.)	3	2	1	1	[25[a], 28, 51]
Discretization method	6	4	0	0	[23, 25, 29, 35, 52, 53]
Discretization levels	12	5	3	2	[25, 29, 33, 36, 37[a], 41, 47[a], 48, 49, 51, 52[a], 54]
Added noise sensitivity	1	0	0	0	[55]
Feature computation					
Texture matrix aggregation	2	1	1	1	[54, 56[a]]
CM symmetry	1	1	1	1	[56[a]]
CM distance	1	1	1	1	[56[a]]
SZM distance	1	1	1	1	[56[a]]
DZM distance	–	–	–	–	–
DZM distance norm	–	–	–	–	–
NGTDM distance	1	1	1	1	[56[a]]
NGLDM distance	–	–	–	–	–
NGLDM coarseness	–	–	–	–	–
Radiomics software	1	1	0	0	[57]
Complementarity	12	8	0	0	[20, 23, 25, 29, 30, 36, 42, 47–49, 53, 54]

CM co-occurrence matrix, *DZM* distance zone matrix, *NGLDM* neighbouring grey level dependence matrix, *NGTDM* neighbourhood grey tone difference matrix, *SZM* size zone matrix, *rec.* reconstructed, *intp.* interpolated

[a] Studies included in the meta-analysis

a factor or image biomarker was absent, weak (o), moderate (+) or strong (++).

Figure 2a shows the sensitivity caused by variation in factors included in the meta-analysis. Of all 38 factors assessed, 31 were evaluated in at least one study and data from 21 were included in the meta-analysis. The level of evidence for factors in the meta-analysis was considered weak due to the low number of studies, with the single exception of test–retest repeatability, for which evidence was deemed moderate. Values for the coefficients on both the logarithmic and original scales and the corresponding 95% credibility intervals are shown in Supplementary Tables 5–7.

Figure 2b–d shows the inherent sensitivity of image biomarkers. All 174 image biomarkers defined in the IBSI reference manual [58] were assessed, and 110 were included in the meta-analysis. Values for the coefficients on both the

logarithmic and original scales and the corresponding 95% credibility intervals are found in Supplementary Tables 8–10.

Acquisition factors

Scan duration was the only acquisition factor with data in the meta-analysis. Variation in scan duration had a non-negligible influence on agreement (PV 1.89, 1.05–4.12, quality of evidence o; CV 2.93, 2.06–4.18, quality of evidence o). However, the range of scan durations assessed varied between studies: 1 and 3 min [21] and 2 and 6 min [22] per bed position. The question as to whether there is a recommended minimum scan duration to ensure overall reproducibility of image biomarkers remains unanswered, but note the relevant guidelines [5].

Both Lovat et al. and Manabe et al. assessed the influence of variations in tracer uptake time [19, 20]. Lovat et al. compared scans with uptake times of 101.5 ± 15.0 and 251.7 ± 18.4 min, whereas Manabe et al. compared scans with uptake times of 60 min and 75 min, but for image biomarkers derived from a polar map. The findings from the two studies are therefore not applicable to typical ^{18}F-FDG PET tumour imaging studies, for which an uptake time of 60 min is recommended, with a range of 55–75 min [5]. Neither study was included in the meta-analysis.

Test–retest repeatability

In test–retest experiments, a human subject or phantom is scanned twice within a time interval of minutes to days. Each scan uses the same acquisition and reconstruction protocol. Interestingly, test–retest affects agreement between biomarker values from both scans more than their reliability (PV 3.80, 3.03–4.8, quality of evidence +; ICC 1.018, 1.001–1.044, quality of evidence +). This indicates that even though variability in differences between biomarker values between scans is substantial, this variability is small compared to intersubject variability.

Reconstruction factors

The evidence for the effect of reconstruction parameters is weak. Variation in the number of subsets for iterative reconstruction appears to be unimportant (CV 1.08, 1.00–1.32, quality of evidence o), but variation in the choice of reconstruction method (PV 2.86, 2.12–3.93, quality of evidence o; CV 2.30, 1.90–2.84, quality of evidence o), number of iterations (CV 1.81, 1.49–2.22, quality of evidence o) and width of the Gaussian filter used for postreconstruction smoothing (CV 2.23, 1.83–2.75, quality of evidence o) do affect agreement. The strongest effect was produced by variations in matrix size (CV 3.63, 2.99–4.47, quality of evidence o). The effect of reconstruction factors on measurement reliability was only determined for variations in reconstruction method, for which it appears to be an important factor as well (ICC 1.071, 1.004–1.273, quality of evidence o). Ideally, reconstruction parameters are kept the same for all imaging data in a radiomics study. In particular, one should be cautious when conducting a radiomics study where PET imaging with and without point spread function (PSF) modelling are mixed. Lasnon et al. indicated that applying a Gaussian filter to a PSF-reconstructed image may reduce differences in such cases [39].

Segmentation factors

All radiomics image biomarkers are computed from a volume of interest that is defined on the PET image. Aside from manual segmentation, a number of autosegmentation methods have been devised that allow either semiautomated or fully automated delineation [59]. According to the meta-analysis, variation in delineation methods affected measurement agreement (PV 2.36, 1.85–3.09, quality of evidence o; CV 2.92, 2.31–3.71, quality of evidence o) and to a lesser extent reliability (ICC 1.033, 1.004–1.076, quality of evidence o). This is likely due to the fact that the segmentation produced differed between the methods [60–63]. To avoid this effect, it is recommended that a single method be used consistently across the entire dataset.

Ideally, the chosen segmentation method is both accurate and robust. Manual segmentation for radiomics should be avoided when possible, as delineation variability between different experts can have a significant effect on agreement between measurements (PV 3.09, 2.33–4.10, quality of evidence o). However, the effect of delineation variability on measurement reliability appeared to be insignificant (ICC 1.003, 1.000–1.015, quality of evidence o). In a recent MICCAI challenge, a number of teams tested automated segmentation algorithms [64]. A convolutional neural network-based method was found to outperform the other methods. Threshold-based techniques were found to perform comparatively poorly. However, Bashir et al. found that semiautomatic threshold-based methods lead to the best reproducibility of image biomarkers when used by different observers [44].

Harmonizing voxel sizes

Differences in voxel size substantially affected measurement agreement (CV 3.63, 2.99–4.47, quality of evidence o). There are two ways to reduce this effect, namely through reconstruction to the same voxel dimensions or by image interpolation. In most situations, re-reconstruction is not possible due to the lack of access to raw data, but image interpolation is possible. Harmonizing voxel sizes was assessed in three studies. Carles et al. compared 3D PET images with 4D PET and found that agreement improved through both reconstruction and image interpolation [25]. Reconstruction reduced fixed biases in

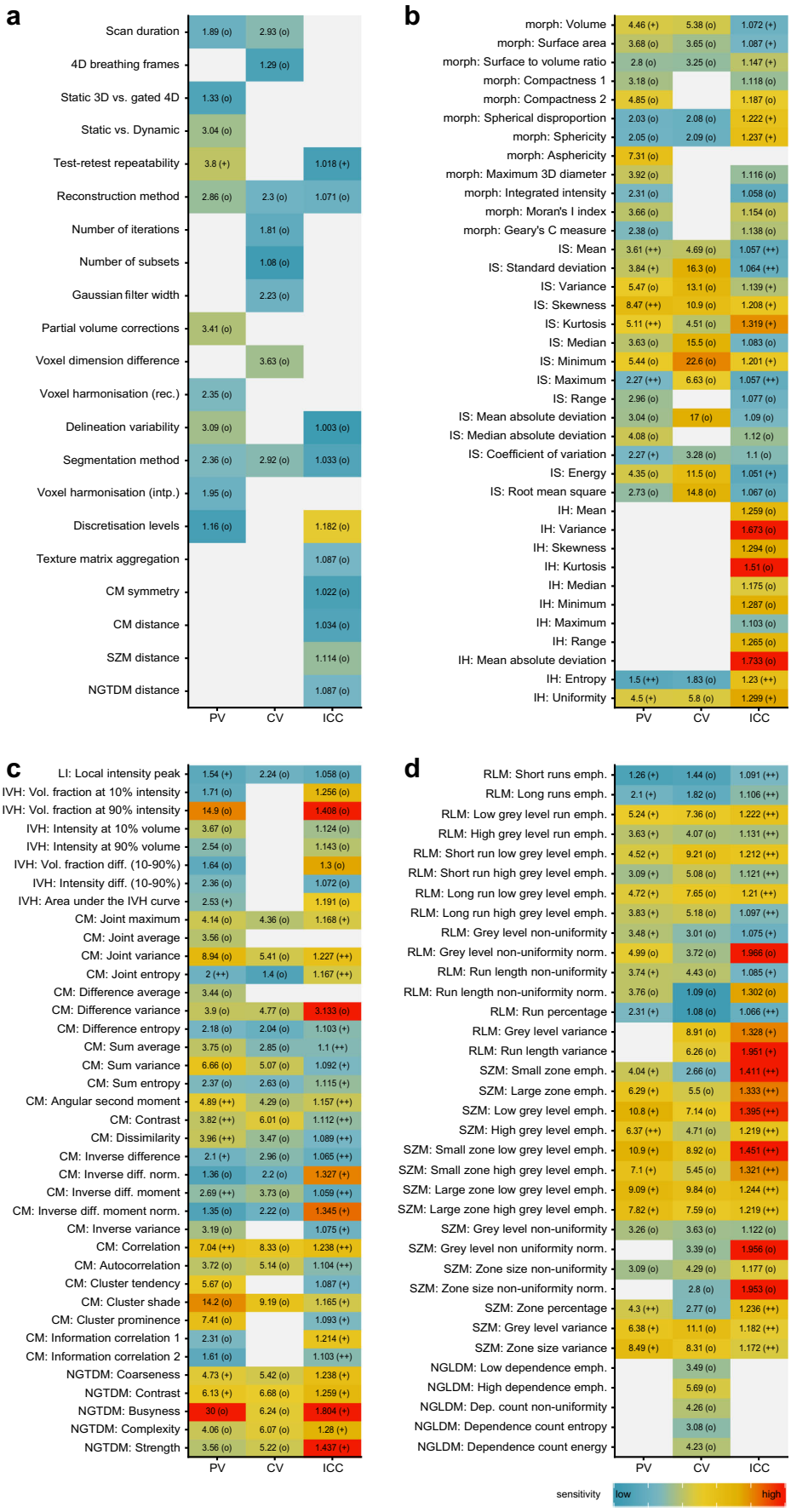

Fig. 2 Sensitivity to variance in factors associated with imaging and radiomics workflow (**a**), and inherent sensitivity of image biomarkers (**b**–**d**). Proportional variability (*PV*) and within-subject coefficient of variation (*CV*) are measures of agreement between image biomarkers, whereas the intraclass correlation coefficient (*ICC*) is a measure of reliability. Sensitivity was estimated on a logarithmic scale from a linear fit with the factor and the image biomarker as predictor variables and the transformed metric as response variable. For comparability of the sensitivity scale between metrics, all estimated coefficients were standardized by dividing by the mean coefficient value of a subset of nine image biomarkers (Supplementary note 1). For each factor and image biomarker, the estimated coefficient is shown on the original metric scale (as a multiplier) together with the quality of evidence (*o* weak, + moderate, ++ strong). IBSI nomenclature is used for image biomarkers: *CM* grey level co-occurrence matrix, *IS*/*IH* intensity-based statistics and intensity histogram, *IVH* intensity-volume histogram, *LI* local intensity, *morph* morphology, *NGLDM* neighbouring grey level dependence matrix, *NGTDM* neighbourhood grey tone difference matrix, *RLM* grey level run length matrix, *SZM* grey level size zone matrix, *rec.* reconstructed, *intp.* interpolated

particular. Reuzé et al. assessed harmonization through interpolation based on comparisons between two cohorts with different scanners and different acquisition and reconstruction protocols using a statistical test, which complicates interpretation of the findings [28]. Yip et al. did not perform a comparison with unharmonized settings [51]. Therefore the extent to which voxel size harmonization increases agreement and reliability cannot be stated with certainty.

The influence of variation in voxel size has been studied more extensively in CT imaging [65–67], where voxel size harmonization was found to be beneficial.

Respiratory motion

Volumes of interest located in the thoracic region may be sensitive to respiratory motion. Gated (4D) PET is used to limit the effect of motion blurring. Current evidence suggests that tumour motion has limited effects on measurement agreement (PV 1.33, 1.03–1.92, quality of evidence o), which was also the conclusion reached by Lovinfosse et al. [7]. The difference in image biomarker values between frames of 4D imaging is also probably small (CV 1.29, 1.01–2.26, quality of evidence o).

Discretization

Discretization is used to decrease the sparsity of the intensity histogram, and consequently that of texture matrices. Fixed bin number and fixed bin size discretization are the most commonly used methods. There are other methods such as Max–Lloyd discretization and histogram equalization, but these are rarely used in PET radiomics. Fixed bin number discretization (relative discretization in reference [8]) is based on the range of SUV intensities found in the volume of interest, with low SUVs corresponding to low bin numbers and high SUVs corresponding to high bin numbers. Hence bin width (in SUV) and SUV range may vary between images in a cohort, even though the number of bins is consistent. Fixed bin size discretization (absolute discretization in reference [8]) creates regularly spaced bins. Hence, the bin width is consistent between images, but the number of bins in the volume of interest and their offset may vary. Both methods have a discretization level parameter, i.e. the number of bins for the fixed bin number method and the bin width in SUV for fixed bin size. Variability due to different discretization levels has a considerable influence on reliability (ICC 1.182, 1.143–1.224, quality of evidence o).

As both methods have their particular advantages and disadvantages [52, 53], they should not be treated as equivalent. The level of discretization can, moreover, determine complementarity with, for example, volume [53, 54] and maximum SUV [53]. Hence, an image biomarker should be considered to consist of the discretization method and level, in addition to its scale and base feature definition [58].

Texture parameters

Many image biomarkers are determined from texture matrices, which capture spatial relationships between neighbouring voxels. All texture matrices require one or more specific parameters for computation. The effect of variation in texture parameters on reproducibility of image biomarkers in PET was only extensively reported by Lv et al. [56]. Some texture parameters are important, such as size zone matrix distance (ICC 1.114, 1.016–1.227, quality of evidence o), texture matrix aggregation (ICC 1.087, 1.026–1.156, quality of evidence o) and neighbourhood grey tone difference matrix distance (ICC 1.087, 1.005–1.273, quality of evidence o). Others were found to be generally unimportant; for example, co-occurrence matrix distance (ICC 1.034, 1.002–1.105, quality of evidence o) and symmetry (ICC 1.022, 1.001–1.079, quality of evidence o). Variability in such parameters is easy to avoid, but requires proper reporting.

Software

Different software tools can produce noticeably different values for the same image biomarker [57, 68]. Minimizing such differences is one of the objectives of the IBSI. At the start of the IBSI, participating teams only managed to produce consensus values for <25% of the image biomarkers, but managed to improve consensus over time [69, 70]. Hence, it is expected that variance due to software will no longer be important after image biomarkers have been standardized.

Image biomarkers

The inherent sensitivity of image biomarkers generally exceeded the sensitivity caused by variation in the various factors assessed above. Of the assessed image biomarkers that had a strong (++) quality of evidence, the most reliable were mean_{IS} (SUV_{mean}) and $\text{maximum}_{\text{IS}}$ (SUV_{max}) with ICCs of 1.057 (1.003–1.162) and 1.057 (1.003–1.159), respectively. The least reliable biomarkers with strong evidence were small $\text{zone emphasis}_{\text{SZM}}$ and small $\text{zone low grey level emphasis}_{\text{SZM}}$ with ICCs of 1.411 (1.206–1.662) and 1.451 (1.240–1.712), respectively. When image biomarkers with moderate (+) quality of evidence were included, the set of the most reliable biomarkers included $\text{energy}_{\text{IS}}$ (ICC 1.051, 1.002–1.180), followed by mean_{IS} and $\text{maximum}_{\text{IS}}$. Inclusion of biomarkers with moderate quality of evidence added $\text{busyness}_{\text{NGTDM}}$ (ICC 1.804, 1.335–2.488) and $\text{run length variance}_{\text{RLM}}$ (ICC 1.951, 1.427–2.687) to the set of least reliable biomarkers mentioned above.

The most reliable image biomarkers were not necessarily those that had the best agreement between measurements. Ranking image biomarkers with moderate or strong quality of evidence according to agreement metrics, $mean_{IS}$ and $maximum_{IS}$ were placed 14th and 7th, respectively, of 42 biomarkers. The evidence for agreement of $energy_{IS}$ was weak.

The expected values for agreement of $mean_{IS}$ and $maximum_{IS}$ under test–retest conditions were similar to those found in the meta-analysis performed by Lodge [3], with PV 12.6% (7.0–22.1%) vs. 14.1 (5.6–22.6%) for $mean_{IS}$ and PV 7.6% (3.5–15.3%) vs. 15.5% (6.3–24.7%) for $maximum_{IS}$. Some reviews have summarized sensitivity according to biomarker family [7, 9]. The current meta-analysis suggests that this approach ignores considerable differences between image biomarkers within a single family. For example, $mean_{IS}$ and $kurtosis_{IS}$ belong to the same family, but differ in agreement and reliability. The issue of nomenclature has been brought up previously [6]. The IBSI has standardized naming of biomarkers and devised a nomenclature system which should help resolve this issue [58].

Data analysis for reproducible radiomics

Image biomarkers are used to assess an outcome of interest. In some cases, this assessment is straightforward. For example, the presence of abnormal ^{18}F-FDG uptake in lymph nodes may indicate nodal spread of cancer. However, in most radiomics studies the assessment is more complex and requires the application of machine learning methods to create statistical models. Machine learning is primarily used to select the information that is important for the problem at hand (feature selection) and to capture this information in a model that allows prediction from new data (model training). The algorithms that perform the latter step are called learners.

The main objective of modelling is to create models that are both relevant and predict well from multiple new datasets. The latter aspect is called generalizability [71]. Generalizable models can be reliably used in different centres with comparable patient populations. Thus, assessing generalizability is important for providing an indication of the external reproducibility of the results, and is usually done by comparing the predictive performance of models between development and validation datasets. A development dataset contains the data that are used to create the model, whereas a validation dataset contains data that are new and were not used during development. Good generalizability exists when model performance is comparable between the development and validation datasets.

Model validation can be internal or external [72]. In internal validation, the dataset is resampled using bootstrapping or cross-validation techniques. These methods split the available data into a development set and a validation set [73]. In external validation, one or more entirely separate datasets are used, which for example come from different clinical centres or studies. A special case of external validation omits the development dataset entirely, and instead validates an externally developed model. External validation gives a stronger indication of generalizability than internal validation [72]. An overview of data analysis strategies is shown in Fig. 3a.

A short introduction to modelling

Many radiomics studies specifically involve supervised learning, where the outcome of interest is available during model development. Some background with regard to the different steps involved in modelling is required before addressing the pitfalls that affect the soundness of the results, generalizability and reproducibility. For more detailed information, textbooks such as Hastie et al. [74] and James et al. [73] are recommended. An overview of modelling steps is shown in Fig. 3b.

Modelling step 1: data preparation

In supervised learning, data consist of predictor variables (features) and a response variable. Sample/subject and batch/cohort identifiers may also be present. In radiomics, predictor variables are image biomarkers and the response variable corresponds to the outcome of interest. The outcome of interest is typically either a categorical variable (e.g. disease yes or no, treatment response or tumour grade good or poor), or a time-to-event variable (e.g. overall survival, disease-free survival) that combines time with an event status.

Data preparation is required to facilitate processing and model development, and may include steps such as transformation, normalization and missing value imputation [75]. Methods used during model development may assume that features follow a normal distribution. Thus, a feature may require transformation to fulfil this assumption. A numerical feature – which most radiomics features are – can often be transformed to follow a normal distribution. Common transformations are Box–Cox power transformation [76] or the more recent Yeo–Johnson power transformation [77]. Whereas transformations affect the distribution of feature values, normalization affects the value range. Many learners require a standardized value range to function properly, which can be obtained through normalization. Typical methods include standardization, which centres values on 0 by subtracting the mean and scales the range by dividing by the standard deviation, and rescaling, which constrains feature values to [0,1] or [−1,1] by dividing a feature by its value range.

Some samples in the data may be missing values for one or more features. Missing values are rarely encountered in radiomics, as features are computed from images, and patients

without imaging are usually excluded. However, the inclusion of clinical data, such as patient age, weight and comorbidities, may introduce incomplete data into the data analysis. Missing values cause issues from both a statistical and an operational point of view. From the statistical point of view, simply omitting all samples with missing data may bias the results [78]. From an operational perspective, missing values are problematic because many computational methods expect complete data and will simply fail to work if data are incomplete. Hence, missing data require imputation, for which various methods have been developed [79, 80].

Data preparation may also include steps that aim to reduce batch effects. In radiomics, batch effects may be caused, for example, through variations in scanners, and acquisition and reconstruction parameters throughout a dataset. Batch effects are likely to be present in multicentre studies, and should be accounted for. Orlhac et al. investigated and validated a batch effect correction method to harmonize multicentre radiomics studies in PET [81, 82]. The suggested ComBat algorithm [83] has recently been used in other radiomics studies for external validation [84, 85]. A simpler algorithm for batch normalization, that is cohort-wise standardization, has also recently been assessed for radiomics [86].

In addition to the above, class imbalances can be addressed during data preparation. Class imbalances can, for example, occur in screening tests when the number of samples that represent a disease state is low compared to samples representing a healthy state. In such situations learners may overemphasize the majority class, which can lead to models that lack sensitivity towards the minority class. Undersampling and oversampling strategies can both be used to address class imbalances [87, 88]. In undersampling, only part of the majority class is used so that classes are balanced, which comes at the cost of removing samples. In oversampling, new samples are generated from the minority class, e.g. using SMOTE [89] or ADASYN [90] algorithms, which incurs the risk of introducing low quality data.

Modelling step 2: data reduction

In radiomics, the number of features usually exceeds the number of samples. Data reduction is used to make processing more efficient and increase learner performance by removing irrelevant and redundant features. Various techniques can be used to perform dimensionality reduction. Some, such as principal component analysis or linear discriminant analysis, project features to a lower dimensional space [91]. The drawback of projection methods is that features are not preserved, and that for reproducibility, all input features need to be available in new data. Other methods assess variable importance and

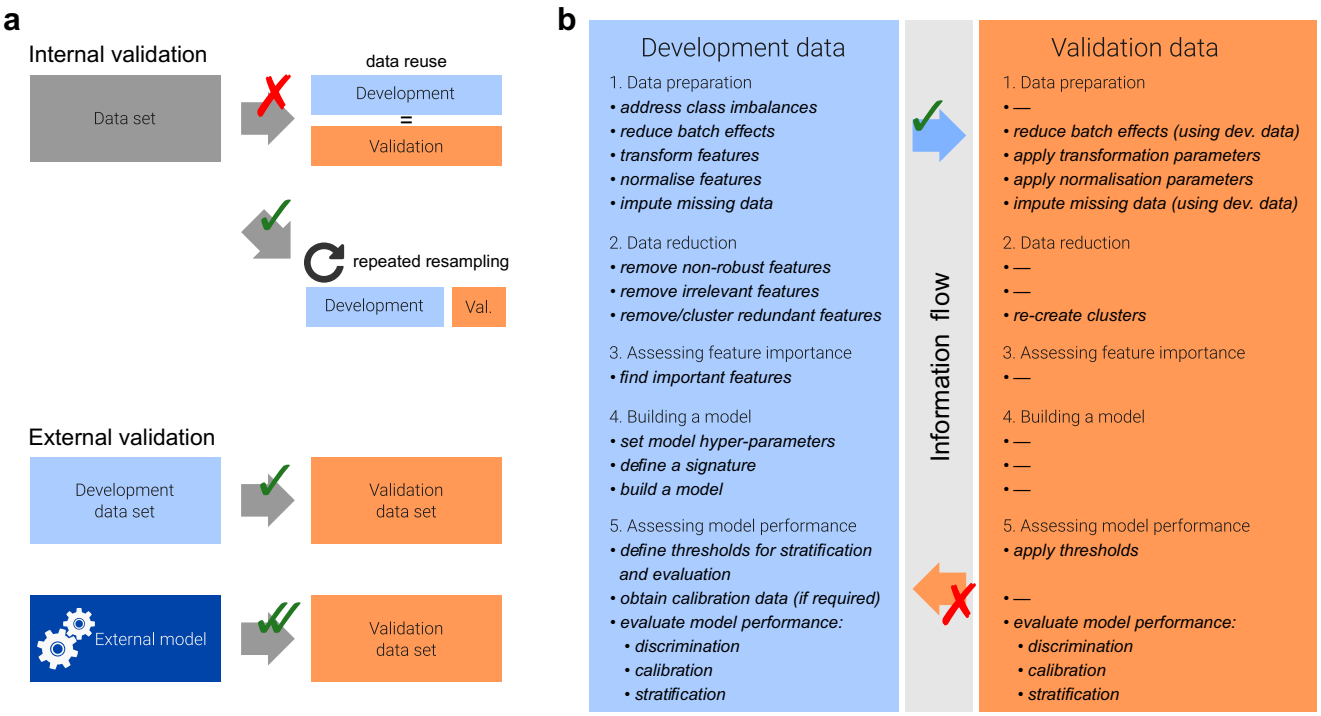

Fig. 3 Data analysis strategies and typical analysis workflow. Generalizability of a radiomics model is assessed through internal and external validation (**a**). Internal validation should be reported by repeatedly dividing the data into development and validation datasets instead of reusing the development data for validation to avoid optimistic biases. The typical workflow involves steps to prepare data, reduce dimensionality, select important features, build a model and evaluate model performance (**b**). Many steps are only fully performed on the development dataset, and the resulting parameters and results are applied to the validation dataset. The aim of this unidirectional information flow is also to avoid optimistic biases

simply eliminate unimportant features from the data [92]. The presence of correlated redundant features may induce a correlation bias [93]. Mutually correlated features may be replaced by a single feature cluster to alleviate the issue [94, 95].

In addition, nonrobust features may be removed [96]. As discussed above, most factors in PET imaging affect feature robustness to a degree. The advantage of removing nonrobust features is that features that are statistically but not causally linked to the outcome of interest may be removed. Removing features can therefore be useful in datasets in which the number of samples is considerably smaller than the number of features and this situation is more likely to occur. However, this may also involve the removal of potentially informative features at the same time, which is a disadvantage. The need to remove nonrobust features thus depends on the particular circumstances. Nonrobust features can be identified using the data presented in Fig. 2. Alternatively, information concerning feature robustness may be obtained through openly available (phantom) datasets, synthetically generated PET data, or through image perturbations [97].

Modelling step 3: assessing feature importance

Models are developed using the most relevant features [98–100]. Feature selection algorithms assess the importance of features. A number of such algorithms have been reviewed in several radiomics studies [95, 101–104], often in conjunction with learners. However, feature selection may be sensitive to perturbations of the dataset [105,106]. To increase the stability of feature selection, it can be performed multiple times using resampled subsets of the development dataset [107–109]. Feature importances in each subset are then subsequently aggregated over all subsets to derive an aggregated feature importance [110]. It should, moreover, be noted that some learners integrate feature selection internally, for example in model-based boosting [111, 112]. For such learners assessing feature importance is not required as an additional step.

Modelling step 4: building a model

The learner uses features and the corresponding outcome to learn the relationship between the two. This information is then encapsulated in a model that can be used to obtain predictions from new data. Hundreds of different learners could potentially be used, depending on the type of outcome [113]. In general, the use of less complex learners such as generalized linear models [114] or the least absolute shrinkage and selection operator [115] is recommended. The resulting models are easy to report and understand, and may serve to establish a performance baseline. Using more complex learners such as random forests [116] and extreme gradient boosting [117] may improve performance further. In such cases, providing the model itself would be required to enable validation by others.

All learners have hyperparameters that require configuration, such as the number of trees in a random forest or the number of features included in a model. Hyperparameters either need to be set manually before any model validation occurs, or require automatic optimization during development to avoid introducing manual biases. Optimization is conducted using the development data and (potentially) a small test set that is separate from the development and validation data. Finding the optimal hyperparameter set for a learner automatically is not trivial. A commonly used method is grid search, which samples the hyperparameter space at regular, predefined positions, and then selects the optimal set of hyperparameters. Grid search is easy to implement and works well if the set of hyperparameters is small. However, grid search is no longer efficient for high-dimensional hyperparameter sets and may actually not find the global optimum. In such cases, random search [118] or more complex methods such as sequential model-based optimization [119] may offer better efficiency and performance.

Modelling step 5: assessing model performance

Model performance should be assessed to indicate how well the model functions and how generalizable it is. For this purpose, the model may be applied to development and validation datasets separately. This produces an expected outcome for every sample. As the true, observed outcome is known for supervised learning, expected and observed outcomes can be compared for accuracy using performance metrics. The suitability of performance metrics depends on the type of outcome. For categorical outcomes, the area under the receiver operating characteristic curve (AUC) is typically measured. The AUC is computed by plotting the true-positive and false-positive rates against one another and calculating the resulting AUC. An AUC score of 1.0 indicates a perfectly discriminating model, whereas a score of 0.5 indicates that the model predictions are essentially random and therefore not meaningful. The AUC also extends to multiple categories [120]. Other useful metrics are clinically relevant metrics such as sensitivity, specificity and accuracy. However, these metrics depend on the balance between outcome classes in the data [121], as well as the chosen class thresholds, and should be used and interpreted carefully. Metrics such as balanced accuracy [122] and Matthews correlation coefficient [123] were devised to avoid dependence on class imbalances.

For time-to-event outcomes, the concordance index (c-index) is commonly measured. The c-index measures the degree of concordance and discordance between sample pairs based on predicted risk or failure time in comparison with observed times and event statuses. The original version was

popularized by Harrell et al. [124]. A version that is less sensitive to censoring was introduced by Uno et al. [125]. The interpretation of c-index values is the same as that for AUC, with 1.0 indicating a perfectly discriminating model and 0.5 a random model. Discrimination on the individual level is not the only relevant aspect of model performance in survival analysis. The ability to stratify patients into two or more risk groups is often clinically relevant and therefore group stratification is also assessed by creating Kaplan–Meier plots and performing log-rank tests [126].

All of the above concerns the discriminative performance of a model. Good discriminative performance, however, does not mean that a method is well-calibrated. Well-calibrated models are clinically important, as they can be used to predict, for example, the expected probability of survival up to 2 years or the expected probability of treatment benefit. Calibration is assessed by comparing the probabilities as observed in the data and the model-based expected probabilities, for example, by determining the intercept and slope of a linear fit [127, 128], as well as through statistical goodness-of-fit tests such as the Hosmer–Lemeshow test [129] and the Nam–D'Agostino test [130, 131].

Modelling pitfalls

The modelling process described above is relatively straightforward. However, many modelling studies have methodological issues that limit model generalizability and validation [132–134]. Such pitfalls can be avoided if they are recognized. The list of common pitfalls was based on modelling experience in our laboratory and issues commonly reported in literature.

Pitfall 1: reusing the development dataset for validation

Many studies measure performance only on the development dataset. This analysis strategy is generally of limited usefulness, as it does not provide an indication of generalizability [135]. The analysis may suggest that a model functions quite well, perhaps even superbly. However, such findings are generally too optimistic and do not stand up to external validation [134]. One important reason is that a model incorporates knowledge concerning data it has seen during development. In the extreme case, a model includes all the knowledge it requires to identify the individual samples from the development dataset [74]. Although such an overfitted model predicts outcomes for the development data very well, the introduction of new and unobserved data will cause predictions to fail. Therefore, generalizability of models should be shown by internal validation using resampling strategies or by external validation [135].

Pitfall 2: using validation data during the model development process

Validation data is used to assess generalizability of a model. As such, validation data should not be used during development and should instead be kept separate at all times [72]. An obvious violation of this principle was mentioned under Pitfall 1. However, the occurrence of validation data leakage may not be evident. For example, in some studies the combined development and validation dataset is inadvertently used for normalization and feature selection, and only the actual modelling and performance assessment steps are performed using separate sets. As feature selection algorithms determine relevance based on the available data, feature selection on a combined dataset will automatically yield features that are relevant for both development and validation data. Such schemes produce biased results when compared with an analysis that does not use validation data for feature selection [135].

Validation data are sometimes also erroneously used when defining thresholds, e.g. for defining classes based on predicted probabilities or for defining risk groups in survival analysis. To avoid positively biased results, such thresholds should be determined using the development data and then applied to the validation dataset.

Pitfall 3: using validation data to optimize the data analysis

Pitfalls 1 and 2 dealt with the use of validation data inside an analysis. However, there are several choices concerning the data analysis itself. Which learner and feature selection method should be used? Which hyperparameter settings will be used, if no automated method is utilized to determine them? Will clustering be used to address feature redundancy? How strictly will irrelevant features be filtered and using what criteria? Such choices can affect performance of the resulting model. The data analysis can be optimized to improve results. This, by itself, is not bad. Different experiments may be required to better understand the data and how to analyse it. However, if the performance on the validation dataset is used to guide the search for a suitable analysis strategy, a positive bias will be introduced into the analysis. It is difficult, if not impossible, to assess whether such bias exists in a reported study. Ideally, a data analysis strategy protocol is defined before performing the analysis. Alternatively, the strategy is fine-tuned using a dataset that is not used for validation afterwards.

Pitfall 4: class imbalances

In the case of categorical outcomes, class imbalances should be addressed. This can be illustrated by the following example. We are interested in stratifying patients for treatment based on the predicted response. In our development data,

10% of the patients have a positive treatment response, and 90% do not. A very naive model would achieve an accuracy of 90% by simply predicting "no response" in all patients. At first glance, the model may appear to be accurate. However, the model and the produced results are misleading and not clinically useful. Class imbalances can be addressed in various ways including data sampling strategies or by using different performance metrics, as discussed earlier. This topic is more fully covered in, for example, references [87, 88].

Pitfall 5: not checking model calibration

Many studies only evaluate the discriminatory performance of models, which measures how well samples can be distinguished from one another. Model calibration is also important to evaluate. Well-calibrated models offer important clinical information, such as the expected disease-free survival probability at 2 years, or treatment success. Such information can be used for treatment stratification. Hence, model calibration should be assessed together with discriminatory performance, as described in the overview of modelling steps.

Pitfall 6: dichotomizing time-to-event outcomes

Time-to-event endpoints are sometimes dichotomized at a certain time point to create two classes, e.g. "tumour recurrence within 2 years" and "no tumour recurrence within 2 years". This approach is sometimes considered to be useful because it enables assessment of survival using tools that may not have implemented methods for processing time-to-event outcomes, or because of clinical interest in the time point. However, it has considerable drawbacks. Dichotomization causes loss of samples if they are censored before the dichotomization time point, information concerning event times is lost, and interpretation of model performance may depend on the choice of dichotomization time point [132, 136, 137]. Dichotomization of time-to-event outcomes should therefore be avoided. If, as in the example above, tumour recurrence within 2 years is of interest, it should be derived from the expected survival probability instead [126].

Pitfall 7: including nonrobust image biomarkers

Including nonrobust image biomarkers in models may lead to models that are poorly generalizable [96]. Interestingly, several radiomics studies have been able to externally validate models without explicitly selecting a subset of robust image biomarkers during model development; see, for example, references [95, 138, 139]. This may indicate that, given sufficient development data, sufficiently robust features are selected for modelling, and noisy nonrobust features are filtered out. Batch normalization, as discussed previously, has also been used to improve reproducibility of features without outright removal of nonrobust features [84, 85]. A robustness analysis has been performed in other studies, and has likewise been able to validate models externally; see, for example, references [140, 141]. Hence, nonrobust biomarkers can be removed, but is not always required.

The robustness of image biomarkers can be determined in several ways. One way is to use the data from the meta-analysis presented in this article to estimate measurement agreement and/or reliability of an image biomarker. However, not all factors could be included, and interaction between factors was not assessed. Another option is to use phantom or simulation measurements and identify nonrobust biomarkers under simplified circumstances. Alternatively, techniques similar to those used for data augmentation in deep learning can be used to estimate image biomarker robustness [97].

Pitfall 8: not comparing results with existing and simple models

A radiomics model is only meaningful if it offers an improvement over an existing (clinical) model, should it exist. Newly developed models should therefore be compared with existing models [6, 142]. Moreover, radiomics models may be difficult to interpret, or use image biomarkers that are not easily explained. It is therefore useful to compare the performance of the suggested radiomics model with that of simple models based on conventional image biomarkers, such as mean and maximum SUV and volume. Similarly, image biomarkers that are included in the model should offer complementary information to conventional biomarkers. A biomarker that is highly correlated with a conventional biomarker can be replaced by the latter to improve interpretability of the model.

Pitfall 9: are the input data correct?

The data used for analysis are often aggregated from multiple sources. One source consists of files with image biomarker values produced by the radiomics software. Another source is formed by outcome data that are commonly found in spreadsheets or tables. These data need to be matched before modelling. Incorrect matching will produce nonsense at best and invalid results in the worst case. Therefore, determining whether matching functioned as expected can prevent issues. Another issue that may occur is that images from the same sample or subject appear in both the development and validation datasets. This should be checked to avoid optimistic bias.

Pitfall 10: incomplete reporting

Adequate reporting is a minimum requirement for reproducing studies and validating results. The critical question that

needs to be answered by an investigator prior to submission is whether the results can be reproduced or validated based on the report and supplementary information. If the answer is no, the missing details should be reported. Reporting guidelines indicate which items should be reported. Radiomics-specific guidelines have been produced by the IBSI [58]. Additionally, the Radiomics Quality Score can help identify weaknesses in the analysis as well as in reporting [96, 143]. Since many radiomics studies involve diagnostic or prognostic modelling, they should be reported according to the "transparent reporting of a multivariable prediction model for individual prognosis or diagnosis" (TRIPOD) guidelines [72, 144]. Other guidelines may be applicable as well; see the EQUATOR network [145].

Ideally, not only should the report be complete, but it should also help other investigators apply the results or methodological improvements. In modelling studies this would mean, for example, that the developed model is provided, accompanied by the parameters required to prepare the input data. The publication of image data and associated metadata moreover broadens the usefulness of the study [146].

Pitfall 11: reinventing the wheel

Radiomics has the advantage of emerging more than a decade after the emergence of the first "omics" fields. This means that mature, reliable software has been created for many parts of the modelling process. R [147] and Python (www.python.org) both have a rich data science ecosystem that can be used for radiomics analysis by installing specific packages or libraries, for example scikit-learn [148], pandas [149], auto-sklearn [150], caret [151], and mlr [152]. The same goes for radiomics-specific software, although these packages are still undergoing maturation and most only cover image processing and image biomarker computation. Examples are pyradiomics [153], CERR [154, 155], LIFEx [156], CaPTk [157, 158], and MITK phenotyping [159]. One can therefore save time and resources by using existing software solutions, and simultaneously minimize the risk of programming errors.

Discussion

The ability to reproduce and validate radiomics studies is vital to generating sufficient and convincing scientific evidence for translating potential applications into clinical practice. In this article we therefore review the reproducibility of radiomics biomarkers in PET imaging and examine radiomics data analysis and several of its pitfalls.

Evidence for sensitivity due to variation in various factors in the meta-analysis was generally weak. There are several reasons for this. One reason is that only studies that assessed agreement using Bland–Altman PV or within-subject coefficient of variation, and studies that assessed reliability using ICC, were included. Moreover, only those studies that used either data recorded in humans or anthropomorphic phantoms were eligible to avoid including results from potentially unrealistic situations. This limited the number of studies eligible for meta-analysis. Second, several eligible studies did not assess all factors using the above-mentioned metrics. Third, data were presented in abstract form in several studies. The authors of these studies were contacted to supply data, but not all replied, and their findings could not be included. More evidence is therefore required to quantify the influence of variation in most factors. This evidence should be reported with one or more of the above metrics in a manner that allows easy integration in a meta-analysis, i.e. in tabular format. The quality of reporting of these studies may, moreover, benefit from following the "Guidelines for Reporting Reliability and Agreement Studies" (GRASS) [15].

Because aggregated evidence for variation in factors was mostly weak, the current meta-analysis cannot be used to provide specific guidelines for harmonizing imaging for PET-based radiomics. However, from the perspective of retrospective or prospective evaluation of imaging data in multicentre studies, it may be prudent to ensure that applicable guidelines for quantitative PET imaging are followed; see, for example, references [5, 160]. Variation up to the reconstruction step should be minimized. Variation in factors that occur after reconstruction can usually and should be avoided. This includes, for example, a consistent choice for the image segmentation algorithm, and using the same discretization scheme for all data.

In comparison with evidence for factors, evidence for image biomarkers was of better quality. However, there are several gaps in the evidence, for example with regard to biomarkers calculated from the distance zone matrix, the intensity histogram and the neighbouring grey level dependence matrix. These biomarker families were rarely or never implemented. Moreover, most studies did not distinguish between biomarkers based on the intensity histogram and statistical biomarkers. According to IBSI definitions these are conceptually different [58]. For example, the $\text{entropy}_{\text{IH}}$ biomarker may have been treated as a statistical feature, instead of an intensity histogram feature, by several studies, in which case it would approximate the logarithm of the number of voxels in the volume of interest.

The meta-analysis shows that for PET imaging, many size zone matrix and neighbourhood grey tone difference biomarkers have a high inherent sensitivity to variation and should be used with care. Conventional biomarkers such as mean_{IS} (SUV_{mean}) and $\text{maximum}_{\text{IS}}$ (SUV_{max}) were found to be reliable.

Aside from evaluating evidence for reproducibility of image biomarkers, several data analysis pitfalls are also described in this article. Avoiding these pitfalls is essential for

creating radiomics models that can be meaningfully used and validated. Particular attention should be paid to ensuring proper assessment of model performance by using a data analysis strategy that involves internal or external validation. Development and validation data should be kept apart at all times. In addition, reporting should be done thoroughly, so that all necessary details for study reproducibility and validation are present.

So far, we have considered radiomics for image biomarkers that are based on hand-crafted features, as this is the only branch of radiomics that has been assessed with regard to reproducibility. A complementary branch of radiomics applies deep learning methods to image analysis. In deep learning, predefined image biomarkers are replaced by features without prior definition that are instead iteratively adapted to fit the development data and integrated in a complex neural network [161, 162]. From the perspective of reproducibility, deep learning may offer several advantages over radiomics using hand-crafted features. First, deep learning replaces the image biomarker computation and modelling steps completely. Second, deep learning can also replace segmentation [163–165] and reconstruction [166] steps. Consequently, factors associated with replaced steps can no longer affect the analysis.

However, deep learning is probably no panacea for resolving reproducibility issues. Deep learning models can be remarkably generalizable [167], but factors that affect reproducibility of image biomarkers are likely also to affect generalizability of deep learning models up to the point where the deep learning model is generated. Moreover, deep learning does not allow the preselection of generally robust image biomarkers. Some model robustness may be induced, either because the development dataset contains the expected heterogeneity, or through data augmentation [168].

It is clear that reproducibility in the context of deep learning-based radiomics needs to be investigated more thoroughly. A possible hindrance here is that reproducibility for deep learning needs to be assessed at the level of the model instead of at the level of the biomarker. This requires that an outcome of interest is obtained in addition to imaging.

In conclusion, we performed a meta-analysis of literature that focused on reproducibility of PET-based image biomarkers and provide an overview of possible data analysis pitfalls. Variations in acquisition, reconstruction, segmentation, radiomics processing and other factors were found to affect agreement and reliability of image biomarkers. Many image biomarkers were, moreover, found to possess a substantial inherent sensitivity to variations.

Acknowledgments The author thanks Dr Jianhua Yan, Dr Matteo Interlenghi, Dr Francesca Gallivanone, Dr Isabella Castiglioni and Dr Lijun Lu for providing data from their studies for use in the meta-analysis.

Compliance with ethical standards

Conflicts of interest None.

Ethical approval This article does not describe any studies with human participants performed by the author.

Informed consent This article describes a meta-analysis on completely anonymous, population-level metrics, and no informed consent was required.

References

1. Kessler LG, Barnhart HX, Buckler AJ, Choudhury KR, Kondratovich MV, Toledano A, et al. The emerging science of quantitative imaging biomarkers terminology and definitions for scientific studies and regulatory submissions. Stat Methods Med Res. 2015;24:9–26.
2. Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible radiomics research for faster clinical translation. J Nucl Med. 2018;59:189–93.
3. Lodge MA. Repeatability of SUV in oncologic 18F-FDG PET. J Nucl Med. 2017;58:523–32.
4. Boellaard R. Standards for PET image acquisition and quantitative data analysis. J Nucl Med. 2009;50(Suppl 1):11S–20S.
5. Boellaard R, Delgado-Bolton R, Oyen WJG, Giammarile F, Tatsch K, Eschner W, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: version 2.0. Eur J Nucl Med Mol Imaging. 2015;42:328–54.
6. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present… any future? Eur J Nucl Med Mol Imaging. 2017;44:151–65.
7. Lovinfosse P, Visvikis D, Hustinx R, Hatt M. FDG PET radiomics: a review of the methodological aspects. Clin Transl Imaging. 2018;6:379–91.
8. Reuzé S, Schernberg A, Orlhac F, Sun R, Chargari C, Dercle L, et al. Radiomics in nuclear medicine applied to radiation therapy: methods, pitfalls, and challenges. Int J Radiat Oncol Biol Phys. 2018;102:1117–42.
9. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys. 2018;102:1143–58.
10. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6:e1000097.
11. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. 2009;6:e1000100.
12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986;1:307–10.
13. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods. 1996;1:30–46.
14. de Vet HCW, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006;59:1033–9.
15. Kottner J, Audige L, Brorson S, Donner A, Gajewski BJ, Hróbjartsson A, et al. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. Int J Nurs Stud. 2011;48:661–71.

16. Nyflot M, Bowen SR, Yang F, Byrd D, Sandison GA, Kinahan PE. Quantitative radiomics: effects of stochastic variability on PET textural features and implications for clinical trials. Int J Radiat Oncol Biol Phys. 2015;93:E566–7.
17. Carles M, Torres-Espallardo I, Alberich-Bayarri A, Olivas C, Bello P, Nestle U, et al. Evaluation of PET texture features with heterogeneous phantoms: complementarity and effect of motion and segmentation method. Phys Med Biol. 2017;62:652–68.
18. 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:e115510.
19. Lovat E, Siddique M, Goh V, Ferner RE, Cook GJR, Warbey VS. The effect of post-injection 18F-FDG PET scanning time on texture analysis of peripheral nerve sheath tumours in neurofibromatosis-1. EJNMMI Res. 2017;7:35.
20. Manabe O, Ohira H, Hirata K, Hayashi S, Naya M, Tsujino I, et al. Use of 18F-FDG PET/CT texture analysis to diagnose cardiac sarcoidosis. Eur J Nucl Med Mol Imaging. 2019;46:1240–7.
21. Bailly C, Bodet-Milin C, Couespel S, Necib H, Kraeber-Bodéré F, Ansquer C, et al. Revisiting the robustness of PET-based textural features in the context of multi-centric trials. PLoS One. 2016;11:e0159984.
22. Grootjans W, Tixier F, van der Vos CS, Vriens D, Le Rest CC, Bussink J, et al. The impact of optimal respiratory gating and image noise on evaluation of intratumor heterogeneity on 18F-FDG PET imaging of lung cancer. J Nucl Med. 2016;57:1692–8.
23. Presotto L, Bettinardi V, De Bernardi E, Belli ML, Cattaneo GM, Broggi S, et al. PET textural features stability and pattern discrimination power for radiomics analysis: an "ad-hoc" phantoms study. Phys Med. 2018;50:66–74.
24. Shiri I, Rahmim A, Ghaffarian P, Geramifar P, Abdollahi H, Bitarafan-Rajabi A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies. Eur Radiol. 2017;27:4498–509.
25. Carles M, Bach T, Torres-Espallardo I, Baltas D, Nestle U, Martí-Bonmatí L. Significance of the impact of motion compensation on the variability of PET image features. Phys Med Biol. 2018;63:065013.
26. Oliver JA, Budzevich M, Zhang GG, Dilling TJ, Latifi K, Moros EG. Variability of image features computed from conventional and respiratory-gated PET/CT images of lung cancer. Transl Oncol. 2015;8:524–34.
27. Tixier F, Vriens D, Cheze-Le Rest C, Hatt M, Disselhorst JA, Oyen WJG, et al. Comparison of tumor uptake heterogeneity characterization between static and parametric 18F-FDG PET images in non-small cell lung cancer. J Nucl Med. 2016;57:1033–9.
28. Reuzé S, Orlhac F, Chargari C, Nioche C, Limkin E, Riet F, et al. Prediction of cervical cancer recurrence using textural features extracted from 18F-FDG PET images acquired with different scanners. Oncotarget. 2017;8:43169–79.
29. Desseroit M-C, Tixier F, Weber WA, Siegel BA, Cheze Le Rest C, Visvikis D, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. J Nucl Med. 2017;58:406–11.
30. Forgacs A, Pall Jonsson H, Dahlbom M, Daver F, DiFranco M, Opposits G, et al. A study on the basic criteria for selecting heterogeneity parameters of F18-FDG PET images. PLoS One. 2016;11:e0164113.
31. Gallivanone F, Interlenghi M, D'Ambrosio D, Trifirò G, Castiglioni I. Parameters influencing PET imaging features: a phantom study with irregular and heterogeneous synthetic lesions. Contrast Media Mol Imaging. 2018;2018:5324517.
32. Leijenaar RTH, Carvalho S, Velazquez ER, van Elmpt WJC, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. Acta Oncol. 2013;52:1391–7.
33. Tixier F, Hatt M, Le Rest CC, Le Pogam A, Corcos L, Visvikis D. Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET. J Nucl Med. 2012;53:693–700.
34. van Velden FHP, Nissen IA, Jongsma F, Velasquez LM, Hayes W, Lammertsma AA, et al. Test-retest variability of various quantitative measures to characterize tracer uptake and/or tracer uptake heterogeneity in metastasized liver for patients with colorectal carcinoma. Mol Imaging Biol. 2014;16:13–8.
35. van Velden FHP, Kramer GM, Frings V, Nissen IA, Mulder ER, de Langen AJ, et al. Repeatability of radiomic features in non-small-cell lung cancer [(18)F]FDG-PET/CT studies: impact of reconstruction and delineation. Mol Imaging Biol. 2016;18:788–95.
36. Willaime JMY, Turkheimer FE, Kenny LM, Aboagye EO. Quantification of intra-tumour cell proliferation heterogeneity using imaging descriptors of 18F fluorothymidine-positron emission tomography. Phys Med Biol. 2013;58:187–203.
37. Altazi BA, Zhang GG, Fernandez DC, Montejo ME, Hunt D, Werner J, et al. Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms. J Appl Clin Med Phys. 2017;18:32–48.
38. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. Acta Oncol. 2010;49:1012–6.
39. Lasnon C, Majdoub M, Lavigne B, Do P, Madelaine J, Visvikis D, et al. 18F-FDG PET/CT heterogeneity quantification through textural features in the era of harmonisation programs: a focus on lung cancer. Eur J Nucl Med Mol Imaging. 2016;43:2324–35.
40. Yan J, Chu-Shern JL, Loi HY, Khor LK, Sinha AK, Quek ST, et al. Impact of image reconstruction settings on texture features 18F-FDG PET. J Nucl Med. 2015;56:1667–73.
41. Doumou G, Siddique M, Tsoumpas C, Goh V, Cook GJ. The precision of textural analysis in (18)F-FDG-PET scans of oesophageal cancer. Eur Radiol. 2015;25:2805–12.
42. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. Eur J Nucl Med Mol Imaging. 2013;40:1662–71.
43. Orlhac F, Nioche C, Soussan M, Buvat I. Understanding changes in tumor textural indices in PET: a comparison between visual assessment and index values in simulated and patient data. J Nucl Med. 2017;58:387–92.
44. Bashir U, Azad G, Siddique MM, Dhillon S, Patel N, Bassett P, et al. The effects of segmentation algorithms on the measurement of 18F-FDG PET texture parameters in non-small cell lung cancer. EJNMMI Res. 2017;7:60.
45. Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell'Oca I, et al. Quantifying the robustness of [18F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. Phys Med. 2018;49:105–11.
46. Takeda K, Takanami K, Shirata Y, Yamamoto T, Takahashi N, Ito K, et al. Clinical utility of texture analysis of 18F-FDG PET/CT in patients with stage I lung cancer treated with stereotactic body radiotherapy. J Radiat Res. 2017;58:862–9.
47. Lu L, Lv W, Jiang J, Ma J, Feng Q, Rahmim A, et al. Robustness of radiomic features in [11C]choline and [18F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. Mol Imaging Biol. 2016;18:935–45.
48. Mu W, Chen Z, Liang Y, Shen W, Yang F, Dai R, et al. Staging of cervical cancer based on tumor heterogeneity characterized by

texture features on (18)F-FDG PET images. Phys Med Biol. 2015;60:5123–39.
49. Orlhac F, Soussan M, Maisonobe J-A, Garcia CA, Vanderlinden B, Buvat I. Tumor texture analysis in 18F-FDG PET: relationships between texture parameters, histogram indices, standardized uptake values, metabolic volumes, and total lesion glycolysis. J Nucl Med. 2014;55:414–22.
50. Wu J, Aguilera T, Shultz D, Gudur M, Rubin DL, Loo BW Jr, et al. Early-stage non-small cell lung cancer: quantitative imaging characteristics of (18)F fluorodeoxyglucose PET/CT allow prediction of distant metastasis. Radiology. 2016;281:270–8.
51. Yip SSF, Parmar C, Kim J, Huynh E, Mak RH, Aerts HJWL. Impact of experimental design on PET radiomics in predicting somatic mutation status. Eur J Radiol. 2017;97:8–15.
52. Leijenaar RTH, Nalbantov G, Carvalho S, van Elmpt WJC, Troost EGC, Boellaard R, et al. The effect of SUV discretization in quantitative FDG-PET radiomics: the need for standardized methodology in tumor texture analysis. Sci Rep. 2015;5:11075.
53. Orlhac F, Soussan M, Chouahnia K, Martinod E, Buvat I. 18F-FDG PET-derived textural indices reflect tissue-specific uptake pattern in non-small cell lung cancer. PLoS One. 2015;10:e0145063.
54. Hatt M, Majdoub M, Vallières M, Tixier F, Le Rest CC, Groheux D, et al. 18F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi-cancer site patient cohort. J Nucl Med. 2015;56:38–44.
55. Oliver JA, Budzevich M, Hunt D, Moros EG, Latifi K, Dilling TJ, et al. Sensitivity of image features to noise in conventional and respiratory-gated PET/CT images of lung cancer: uncorrelated noise effects. Technol Cancer Res Treat. 2017;16:595–608.
56. Lv W, Yuan Q, Wang Q, Ma J, Jiang J, Yang W, et al. Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT. Eur Radiol. 2018;28:3245–54.
57. Bogowicz M, Leijenaar RTH, Tanadini-Lang S, Riesterer O, Pruschy M, Studer G, et al. Post-radiochemotherapy PET radiomics in head and neck cancer – the influence of radiomics implementation on the reproducibility of local control tumor models. Radiother Oncol. 2017;125:385–91.
58. Zwanenburg A, Leger S, Vallières M, Löck S, for the Image Biomarker Standardisation Initiative. Image biomarker standardisation initiative [Internet]. arXiv:1612.07003 [cs.CV]. 2016. http://arxiv.org/abs/1612.07003.
59. Hatt M, Lee JA, Schmidtlein CR, Naqa IE, Caldwell C, De Bernardi E, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM task group no. 211. Med Phys. 2017;44:e1–42.
60. Nestle U, Kremp S, Schaefer-Schuler A, Sebastian-Welsch C, Hellwig D, Rübe C, et al. Comparison of different methods for delineation of 18F-FDG PET-positive tissue for target volume definition in radiotherapy of patients with non-small cell lung cancer. J Nucl Med. 2005;46:1342–8.
61. Schinagl DAX, Vogel WV, Hoffmann AL, van Dalen JA, Oyen WJ, Kaanders JHAM. Comparison of five segmentation tools for 18F-fluoro-deoxy-glucose-positron emission tomography-based target volume definition in head and neck cancer. Int J Radiat Oncol Biol Phys. 2007;69:1282–9.
62. Zaidi H, El Naqa I. PET-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. Eur J Nucl Med Mol Imaging. 2010;37:2165–87.
63. Cheebsumon P, Yaqub M, van Velden FHP, Hoekstra OS, Lammertsma AA, Boellaard R. Impact of [18F]FDG PET imaging parameters on automatic tumour delineation: need for improved tumour delineation methodology. Eur J Nucl Med Mol Imaging. 2011;38:2136–44.
64. Hatt M, Laurent B, Ouahabi A, Fayad H, Tan S, Li L, et al. The first MICCAI challenge on PET tumor segmentation. Med Image Anal. 2018;44:177–95.
65. Shafiq-Ul-Hassan M, Zhang GG, Latifi K, Ullah G, Hunt DC, Balagurunathan Y, et al. Intrinsic dependencies of CT radiomic features on voxel size and number of gray levels. Med Phys. 2017;44:1050–62.
66. Mackin D, Fave X, Zhang L, Yang J, Jones AK, Ng CS, et al. Harmonizing the pixel size in retrospective computed tomography radiomics studies. PLoS One. 2017;12:e0178524.
67. Larue RTHM, van Timmeren JE, de Jong EEC, Feliciani G, Leijenaar RTH, Schreurs WMJ, 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;56:1544–53.
68. Foy JJ, Robinson KR, Li H, Giger ML, Al-Hallaq H, Armato SG. Variation in algorithm implementation across radiomics software. J Med Imaging. 2018;5:044505.
69. Zwanenburg A, Abdalah MA, Apte A, Ashrafinia S, Beukinga J, Bogowicz M, et al. PO-0981: results from the Image Biomarker Standardisation Initiative. Radiother Oncol. 2018;127:S543–4.
70. Hatt M, Vallieres M, Visvikis D, Zwanenburg A. IBSI: an international community radiomics standardization initiative. J Nucl Med. 2018;59(Suppl 1):287–7.
71. Domingos P. A few useful things to know about machine learning. Commun ACM. 2012;55:78–87.
72. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162:W1–73.
73. James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning: with Applications in R. New York: Springer; 2013.
74. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second ed. New York: Springer Science+Business Media; 2009.
75. García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. New York: Springer; 2015.
76. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Ser B Stat Methodol. 1964;26:211–52.
77. Yeo I, Johnson RA. A new family of power transformations to improve normality or symmetry. Biometrika. 2000;87:954–9.
78. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol. 1995;142:1255–64.
79. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59:1087–91.
80. Luengo J, García S, Herrera F. On the choice of the best imputation methods for missing values considering three groups of classification methods. Knowl Inf Syst. 2012;32:77–108.
81. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. J Nucl Med. 2018;59:1321–8.
82. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT radiomics. Radiology. 2019;291:53–9.
83. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
84. Lucia F, Visvikis D, Vallières M, Desseroit M-C, Miranda O, Robin P, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2019;46:864–77.

85. Foley KG, Shi Z, Whybra P, Kalendralis P, Larue R, Berbee M, et al. External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. Radiother Oncol. 2019;133:205–12.
86. Chatterjee A, Vallières M, Dohan A, Levesque IR, Ueno Y, Saif S, et al. Creating robust predictive radiomic models for data from independent institutions using normalization. IEEE TRPMS. 2019;3:210–5.
87. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21:1263–84.
88. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5:221–32.
89. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
90. Haibo He, Yang Bai, Garcia EA, Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). 2008. p. 1322–8.
91. Cunningham JP, Ghahramani Z. Linear dimensionality reduction: survey, insights, and generalizations. J Mach Learn Res. 2015;16:2859–900.
92. John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. In: Cohen WW, Hirsh H, editors. Machine Learning Proceedings 1994. San Francisco: Morgan Kaufmann; 1994. p. 121–9.
93. Tolosi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. Bioinformatics. 2011;27:1986–94.
94. Park MY, Hastie T, Tibshirani R. Averaged gene expressions for regression. Biostatistics. 2007;8:212–27.
95. Leger S, Zwanenburg A, Pilz K, Lohaus F, Linge A, Zöphel K, et al. A comparative study of machine learning methods for time-to-event survival data for radiomics risk modelling. Sci Rep. 2017;7:13206.
96. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14:749–62.
97. Zwanenburg A, Leger S, Agolli L, Pilz K, Troost EGC, Richter C, et al. Assessing robustness of radiomic features by image perturbation. Sci Rep. 2019;9:614.
98. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157–82.
99. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23:2507–17.
100. Li J, Cheng K, Wang S, Morstatter F, Trevino RP, Tang J, et al. Feature selection: a data perspective. ACM Computing Surveys. 2018;50:94.
101. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. Sci Rep. 2015;5:13087.
102. Parmar C, Grossmann P, Rietveld D, Rietbergen MM, Lambin P, Aerts HJWL. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. Front Oncol. 2015;5:272.
103. Zhang B, He X, Ouyang F, Gu D, Dong Y, Zhang L, et al. Radiomic machine-learning classifiers for prognostic biomarkers of advanced nasopharyngeal carcinoma. Cancer Lett. 2017;403:21–7.
104. Sun W, Jiang M, Dang J, Chang P, Yin F-F. Effect of machine learning methods on predicting NSCLC overall survival time based on Radiomics analysis. Radiat Oncol. 2018;13:197.
105. Kalousis A, Prados J, Hilario M. Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl Inf Syst. 2007;12:95–116.
106. Haury A-C, Gestraud P, Vert J-P. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One. 2011;6:e28210.
107. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Daelemans W, Goethals B, Morik K, editors. Machine Learning and Knowledge Discovery in Databases. Berlin: Springer; 2008. p. 313–25.
108. Abeel T, Helleputte T, Van de Peer Y, Dupont P, Saeys Y. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics. 2010;26:392–8.
109. Meinshausen N, Bühlmann P. Stability selection. J R Stat Soc Ser B Stat Methodol. 2010;72:417–73.
110. Wald R, Khoshgoftaar TM, Dittman D, Awada W, Napolitano A. An extensive comparison of feature ranking aggregation techniques in bioinformatics. 2012 IEEE 13th International Conference on Information Reuse Integration (IRI). 2012; p. 377–84.
111. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. Stat Sci. 2007;22:477–505.
112. Hofner B, Boccuto L, Göker M. Controlling false discoveries in high-dimensional situations: boosting with stability selection. BMC Bioinformatics. 2015;16:144.
113. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15:3133–81.
114. Nelder JA, Wedderburn RWM. Generalized linear models. J R Stat Soc Ser A. 1972;135:370–84.
115. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Stat Methodol. 1996;58:267–88.
116. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
117. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–94.
118. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281–305.
119. Hutter F, Hoos HH, Leyton-Brown K. Sequential model-based optimization for general algorithm configuration. In: Coello CAC, editor. Learning and Intelligent Optimization. Berlin: Springer; 2011. p. 507–23.
120. Hand DJ, Till RJ. A simple generalisation of the area under the ROC curve for multiple class classification problems. Mach Learn. 2001;45:171–86.
121. Japkowicz N, Stephen S. The class imbalance problem: a systematic study. Intell Data Anal. 2002;6:429–49.
122. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. 20th International Conference on Pattern Recognition. 2010; p. 3121–4.
123. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 2000;16:412–24.
124. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 1996;15:361–87.
125. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei LJ. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Stat Med. 2011;30:1105–17.
126. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Methodol. 2013;13:33.
127. Miller ME, Langefeld CD, Tierney WM, Hui SL, McDonald CJ. Validation of probabilistic predictions. Med Decis Mak. 1993;13:49–58.
128. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction

models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38.
129. Hosmer DW Jr, Lemeshow S, Sturdivant RX. Applied Logistic Regression. Hoboken: Wiley; 2013.
130. D'Agostino RB, Nam B-H. Evaluation of the performance of survival analysis models: discrimination and calibration measures. In: Balakrishnan N, Rao CR, editors. Handbook of Statistics. Amsterdam: Elsevier; 2003; p. 1–25.
131. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. Stat Med. 2015;34:1659–80.
132. Dupuy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. J Natl Cancer Inst. 2007;99:147–57.
133. Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, et al. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol. 2010;28:827–38.
134. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: a systematic review. PLoS One. 2015;10:e0124165.
135. Zwanenburg A, Löck S. Why validation of prognostic models matters? Radiother Oncol. 2018;127:370–3.
136. Binder H, Porzelius C, Schumacher M. An overview of techniques for linking high-dimensional molecular data to time-to-event endpoints by risk prediction models. Biom J. 2011;53:170–89.
137. Chen H-C, Kodell RL, Cheng KF, Chen JJ. Assessment of performance of survival prediction models for cancer prognosis. BMC Med Res Methodol. 2012;12:102.
138. Vallières M, Kay-Rivest E, Perrin LJ, Liem X, Furstoss C, Aerts HJWL, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Sci Rep. 2017;7:10117.
139. Sun R, Limkin EJ, Vakalopoulou M, Dercle L, Champiat S, Han SR, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. Lancet Oncol. 2018;19:1180–91.
140. Aerts HJWL, Velazquez ER, Leijenaar RTH, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006.
141. Leger S, Zwanenburg A, Pilz K, Zschaeck S, Zöphel K, Kotzerke J, et al. CT imaging during treatment improves radiomic models for patients with locally advanced head and neck cancer. Radiother Oncol. 2019;130:10–7.
142. Welch ML, McIntosh C, Haibe-Kains B, Milosevic MF, Wee L, Dekker A, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol. 2019;130:2–9.
143. Sanduleanu S, Woodruff HC, de Jong EEC, van Timmeren JE, Jochems A, Dubois L, et al. Tracking tumor biology with radiomics: a systematic review utilizing a radiomics quality score. Radiother Oncol. 2018;127:349–60.
144. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg. 2015;102:148–58.
145. Simera I, Moher D, Hirst A, Hoey J, Schulz KF, Altman DG. Transparent and accurate reporting increases reliability, utility, and impact of your research: reporting guidelines and the EQUATOR Network. BMC Med. 2010;8:24.
146. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
147. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2018. https://www.R-project.org/
148. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
149. McKinney W. Data structures for statistical computing in Python. Austin: Proceedings of the 9th Python in Science Conference; 2010. p. 51–6.
150. Feurer M, Klein A, Eggensperger K, Springenberg J, Blum M, Hutter F. Efficient and robust automated machine learning. In: Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R, editors. Advances in Neural Information Processing Systems 28. New York: Curran Associates; 2015. p. 2962–70.
151. Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;28:1–26.
152. Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, et al. mlr: machine learning in R. J Mach Learn Res. 2016;17:5938–42.
153. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational radiomics system to decode the radiographic phenotype. Cancer Res. 2017;77:e104–7.
154. Deasy JO, Blanco AI, Clark VH. CERR: a computational environment for radiotherapy research. Med Phys. 2003;30:979–85.
155. Apte AP, Iyer A, Crispin-Ortuzar M, Pandya R, van Dijk LV, Spezi E, et al. Technical note: extension of CERR for computational radiomics: a comprehensive MATLAB platform for reproducible radiomics research. Med Phys. 2018;45:3713–20.
156. Nioche C, Orlhac F, Boughdad S, Reuzé S, Goya-Outi J, Robert C, et al. LIFEx: a freeware for radiomic feature calculation in multimodality imaging to accelerate advances in the characterization of tumor heterogeneity. Cancer Res. 2018;78:4786–9.
157. Davatzikos C, Rathore S, Bakas S, Pati S, Bergman M, Kalarot R, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. J Med Imaging. 2018;5:011018.
158. Rathore S, Bakas S, Pati S, Akbari H, Kalarot R, Sridharan P, et al. Brain cancer imaging phenomics toolkit (brain-CaPTk): an interactive platform for quantitative analysis of glioblastoma. In: Crimi A, Bakas S, Kuijf H, Menze B, Reyes M, editors. Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Cham: Springer; 2018. p. 133–45.
159. Götz M, Nolden M, Maier-Hein K. MITK phenotyping: an open-source toolchain for image-based personalized medicine with radiomics. Radiother Oncol. 2019;131:108–11.
160. Fendler WP, Eiber M, Beheshti M, Bomanji J, Ceci F, Cho S, et al. 68Ga-PSMA PET/CT: Joint EANM and SNMMI procedure guideline for prostate cancer imaging: version 1.0. Eur J Nucl Med Mol Imaging. 2017;44:1014–24.
161. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
162. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys. 2019;46:e1–36.
163. Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer, LNCS, vol. 9351. 2015. p. 234–41.
164. Milletari F, Navab N, Ahmadi S. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV). IEEE; 2016. p. 565–71.
165. Isensee F, Petersen J, Klein A, Zimmerer D, Jaeger PF, Kohl S, et al. nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation. arXiv [cs.CV]. 2018. http://arxiv.org/abs/1809.10486.
166. Zhu B, Liu JZ, Cauley SF, Rosen BR, Rosen MS. Image reconstruction by domain-transform manifold learning. Nature. 2018;555:487–92.

167. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. arXiv [cs.LG]. 2016. http://arxiv.org/abs/1611.03530.
168. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal. 2017;42:60–88.