European Journal of Nuclear Medicine and Molecular Imaging (2019) 46:2673–2699
https://doi.org/10.1007/s00259-019-04414-4

REVIEW ARTICLE

AI-based applications in hybrid imaging: how to build smart and truly multi-parametric decision models for radiomics

Isabella Castiglioni[1] · Francesca Gallivanone[1] · Paolo Soda[2] · Michele Avanzo[3] · Joseph Stancanello[3,4] · Marco Aiello[5] · Matteo Interlenghi[1] · Marco Salvatore[5]

Received: 10 June 2019 / Accepted: 21 June 2019 / Published online: 11 July 2019

Abstract

Introduction The quantitative imaging features (radiomics) that can be obtained from the different modalities of current-generation hybrid imaging can give complementary information with regard to the tumour environment, as they measure different morphologic and functional imaging properties. These multi-parametric image descriptors can be combined with artificial intelligence applications into predictive models. It is now the time for hybrid PET/CT and PET/MRI to take the advantage offered by radiomics to assess the added clinical benefit of using multi-parametric models for the personalized diagnosis and prognosis of different disease phenotypes.

Objective The aim of the paper is to provide an overview of current challenges and available solutions to translate radiomics into hybrid PET-CT and PET-MRI imaging for a smart and truly multi-parametric decision model.

Keywords Radiomics · Artificial intelligence · Decision models · Hybrid imaging · PET/CT · PET/MRI

Introduction

Since its advent into clinical practice, medical imaging methods, such as computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET), have acquired a central role in the clinical management of a wide variety of diseases, thanks to the unique possibility of characterizing, in vivo and non-invasively, the presence of physio-pathological processes at different stages of diseases and even following therapies.

In particular, improvements in PET imaging system technologies, including three-dimensional (3D) acquisition and iterative reconstruction as well as availability of new molecular probes, has increased accuracy, sensitivity, specificity, and lesion detectability of PET imaging studies [1].

Hybrid PET/CT scanners were commercially introduced as diagnostic systems in 2001 in order to provide non-invasive imaging assessment of both body function and structure in a single examination. In the almost 20 years since then, these systems have been recognized as the most recommended tool to detect, localize, and characterize many multifactorial diseases. The impact of hybrid PET/CT imaging has emerged, in particular, in oncology, neurology, and cardiology, where it has currently gained a consolidated role in diagnosis, staging, follow-up, and treatment monitoring and planning.

Hybrid PET/MRI systems enteredthe market only recently. MRI is superior to CT in providing images with high soft-tissue contrast and spatial resolution; thus, PET/MRI systems could offer better lesion detectability and characterization than PET/CT with a decreased radiation dose. However, such systems present many technical elements of complexity and costs with

Isabella Castiglioni and Paolo Soda contributed equally to this work.

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

✉ Paolo Soda
p.soda@unicampus.it

[1] Institute of Molecular Imaging and Physiology, National Research Council (IBFM-CNR), 20090 Segrate, MI, Italy

[2] Unit of Computer Systems and Bioinformatics, Department of Engineering, Università Campus Bio-Medico di Roma, 00128 Rome, Italy

[3] Medical Physics, Centro di Riferimento Oncologico di Aviano (CRO) IRCCS, 33081 Aviano, Italy

[4] Radiologia e Diagnostica per Immagini, Centro Diagnostico Italiano, 20147 Milan, Italy

[5] IRCCS SDN, Istituto di Ricerca Diagnostica e Nucleare, Naples, Italy

respect to hybrid PET/CT. Their installation requires appropriate environmental solutions such as shielding the operative rooms from both radiation and magnetic field effects. Furthermore, PET/MRI technology needs complex acquisition protocols and long scan duration, in particular when multi-parametric or whole-body PET/MRI studies are required. Moreover, image quality and quantitation are more difficult to achieve than for PET/CT systems. As an example, contrary to CT, MRI signal does not allow a direct association to density maps to be applied straightforwardly for attenuation correction of PET images, making attenuation correction more difficult [2]. Notwithstanding the speculations that these systems would not grow rapidly in number if clear clinical benefits did not continue to be demonstrated, a number of PET/MRI integrated scanners are currently available worldwide. Most published results are indicating that hybrid PET/MRI used with clinical current practice of each stand-alone imaging modality shows comparable performance in the detection, diagnosis, and location of suspected diseases as compared to hybrid PET/CT and MRI, with the most recognized added value of a unique examination session.

In clinical current practice of hybrid PET imaging, mainly aimed at diagnosis, staging, and treatment monitoring, at a first level, PET images are evaluated qualitatively. The CT component is used to better localize and characterize lesions identified on the PET images. The MRI images are, in turn, multimodal and can support PET images not only by supplying anatomical reference, but also by integrating functional characterization of the lesion, for example with microstructural features by diffusion weighted imaging (DWI) and perfusion parameters by dynamic contrast enhanced (DCE) MRI.

Diseases are then described and reported by nuclear medicine physicians and radiologists using semantic lexicon features, e.g., related to the presence of disease aggressive behavior, infiltration and metastatic capacity, and likelihood of recurrence, even following therapy. However, qualitative analysis of medical images presents several limitations, including the subjective expertise of the referring physicians in image interpretation, and their limited human ability to capture subtle disease features and their temporal changes using the naked eye.

In recent years, a great effort has been devoted to overcoming these limitations. Quantitative approaches to medical image analysis have been developed and implemented thanks to compensation for the physical limitations of imaging systems, e.g., poor spatial resolution in the PET images from hybrid PET/CT [e.g. 3–6]. The technological evolution of hybrid imaging is beyond the purpose of this review, as well as the comparison of PET/CT and PET/MRI, since the literature is full of extensive works on this topic. However, it is worth noting how such systems started to lead to an enormous amount of data generated by the two different image components of a hybrid system, e.g., the PET, low-dose CT, and contrast-enhanced CT from hybrid PET/CT, the PET and MRI (in multiple modal acquisitions) from hybrid PET/MRI.

Both spatial and temporal intra-tumor heterogeneity have been recognized as important biological characteristics impacting on resistance to therapy and evolution of cancer diseases. Since genomic characterization of tumor by ex-vivo tumor biopsy showed clear limitations due to the regional dependence of results from the tumor sample region [7], the amount of complementary, multi-modal, and multi-parametric imaging data, co-registered in spatial distribution and time, offered unique insights into opportunities to explore intra-tumor heterogeneity at the macroscopic scale in both the anatomical and functional dimensions, in vivo and non invasively.

With this aim, advanced image processing methods have been developed for extracting quantitative “features” from the images of an entire tumor, capturing such heterogeneity as the cause of different clinical outcomes for patients with similar diagnosis or following similar therapies. Such methods, as for example image texture analysis, capturing appearance, structure, and arrangement of the tumor in an image, were first applied at a limited scale, that is considering a few imaging features from lesions to better characterize the tumour environment. However, these methods allow a very large amount of quantitative data to be obtained for a single patient, thus now configuring medical imaging as a source of big data and posing the challenge of evaluating how to obtain meaningful information correlating such data with clinical and genomic data to personalize the prognosis and treatment of the single patient. The new paradigm, “radiomics”, is then emerging as the high-throughput extraction of quantitative features from medical images to characterize phenotype such as the in-vivo expression of the genotype of different diseases, based on the assumption that stratification of patients into subtypes is possible using such radiomic biomarkers [8]. Moreover, artificial intelligence [AI] methods [9] have been applied in radiomics to predict what will be associated with treatment strategies such as tumor subtypes, survival time, and disease recurrence [10]. Predictive models can be built on radiomic multi-parametric image descriptors to personalize the decision-making for patients [11, 12]. Even if still today the radiomic approach has been mainly applied to oncological malignancies, this approach is also emerging in other diseases such as neurological, cerebrovascular, and immunological diseases [e.g. 13–15].

These achievements have led to the investigation of a new clinical role for medical imaging, shifting the focus from the diagnosis to the prognosis of diseases. This paradigm shift has already started for structural imaging such as CT and MRI: their high resolution and contrast have allowed radiologists the direct observation of different image descriptors, providing proof-of-concept of radiomics; their significant use for diagnostic purposes in several oncological studies has allowed the availability of large cohorts of patient images for the selection and validation of radiomic prognostic signatures [16].

In PET, the first study including radiomic analysis was published in 2009 [17]. The majority of the subsequent radiomic studies have included less than 100 patients [18], limiting the possibility of developing and assessing predictive models.

With the use of hybrid PET/CT and PET/MRI, the combined modalities should be able to highlight the complementary tumor characteristics in order to maximize the information that can be extracted with radiomics. The aim of our paper is to provide an overview of current challenges and available solutions that can help in translating radiomics in hybrid PET imaging, paying attention to those specific issues needing to be carefully addressed to build a smart and truly multiparametric decision model.

Radiomic workflow in hybrid imaging

Even if the radiomics pipeline was well known in the image processing community, one of the most important issues at the beginning of radiomics was its translation in the context of radiological images and its standardization. A great effort was dedicated to this purpose and many methodological papers have been put together, leading to a shared robust methodological framework on radiomics that consists of several sequential steps to be performed as prerequisites for building decision models [19].

Standardization of radiomics is fundamental to guarantee the reproducibility of the results; that is, the ability of radiomic image processing methodology to return inter- and intra-operator independent feature estimates to be used to build decision models.

Image processing steps can present complexities and be a source of variability when applied to radiomics from hybrid imaging, but at the same time they can benefit from the availability of the different imaging modalities offered by hybrid systems to find solutions that cannot be applied based on a single imaging modality.

Choice of the study protocol and data collection

The choice of the imaging protocol from hybrid PET systems should be guided by the criterion of the best modality able to capture the presence of a certain level of tissue heterogeneity underlying the disease under study, but should also follow considerations concerning the clinical appropriateness for the patient under study, avoiding over-examinations. Intratumor heterogeneity should be recognized to lead to disease subtypes with different prognosis (even following therapy) that could be misdiagnosed by the tumor gene expression profiles as obtained from tumor-biopsy samples on small fractions of the tumor [7].

The radiomic analysis should be applied for those tumors that are expected to show other components than the solid one, usually associated to a better prognosis. However, there are some cases, for example bronchiolo-alveolar carcinoma, where tumors with a high percentage of solid components have a worse prognosis [20] that could benefit from radiomic evaluation.

Radiomic analysis should be properly designed also by establishing in which stage of the disease the radiomic features should be extracted and used, for example at the baseline rather than after first treatments or at the end of the therapy [e.g., 21].

Physiopathological characteristics underlying cancer, such as metabolism, vascularization, perfusion, cell proliferation, necrosis, hypoxia, and gene expression, can be measured by PET with [18F]fluorodeoxyglucose ([18F]FDG) [22–24]. Thus, we can expect that different values of PET radiomic features may be measured in benign and malignant lesions as a consequence of the different underlying physiopathological processes. [18F]FDG is recommended also for radiotherapy evaluation during follow-up and for therapy planning based on PET images, and radiomics have been shown to be able to address different responses to radiotherapy treatments in many oncological studies. Moreover, hypoxia was found to impact the radio-chemotherapy outcome and even surgery outcome. This occurs both for large tumors with necrosis, and for small primary tumors and recurrences, micro-metastases, and surgical margins showing microscopic tumor involvement. Thus, PET hypoxia radiotracers could capture proper characteristics for radiotherapy response [25].

However, PET images have more physical limitations for radiomics than MRI or CT, due to a worse spatial resolution, leading to poorer spatial sampling and lower signal-to-noise ratio. In addition, PET images are often heavily filtered during reconstruction, thus reducing the heterogeneity expression of the image [26]. In hybrid PET studies, radiomic features extracted from the low-dose CT component of PET/CT, usually without contrast enhancement, can measure the heterogeneity of vascularization, necrosis, or cellularity, as well as the proportions of fat, air, and water [27]. Contrast-enhanced CT, used as an additional CT sequence in PET/CT studies, and the MRI component in PET/MRI studies, can measure the heterogeneity of vessel density, perfusion, proton density, and physiological tissue characteristics at the level of tumor physiological habitats [e.g., 24, 28–31]. An example of the variety of image modalities offered by current-generation hybrid PET/MRI is shown in Fig. 1.

It is then fundamental to establish what the target of the radiomic study is. If the study target is the biological characterization, other data should be collected from the patients, e.g., biological or molecular biomarkers or protein/gene expression levels, in order to assess associations of radiomics features with known biological characteristics of disease. If the study target is the prognosis or therapy response, it is important to collect clinical data of patients during the

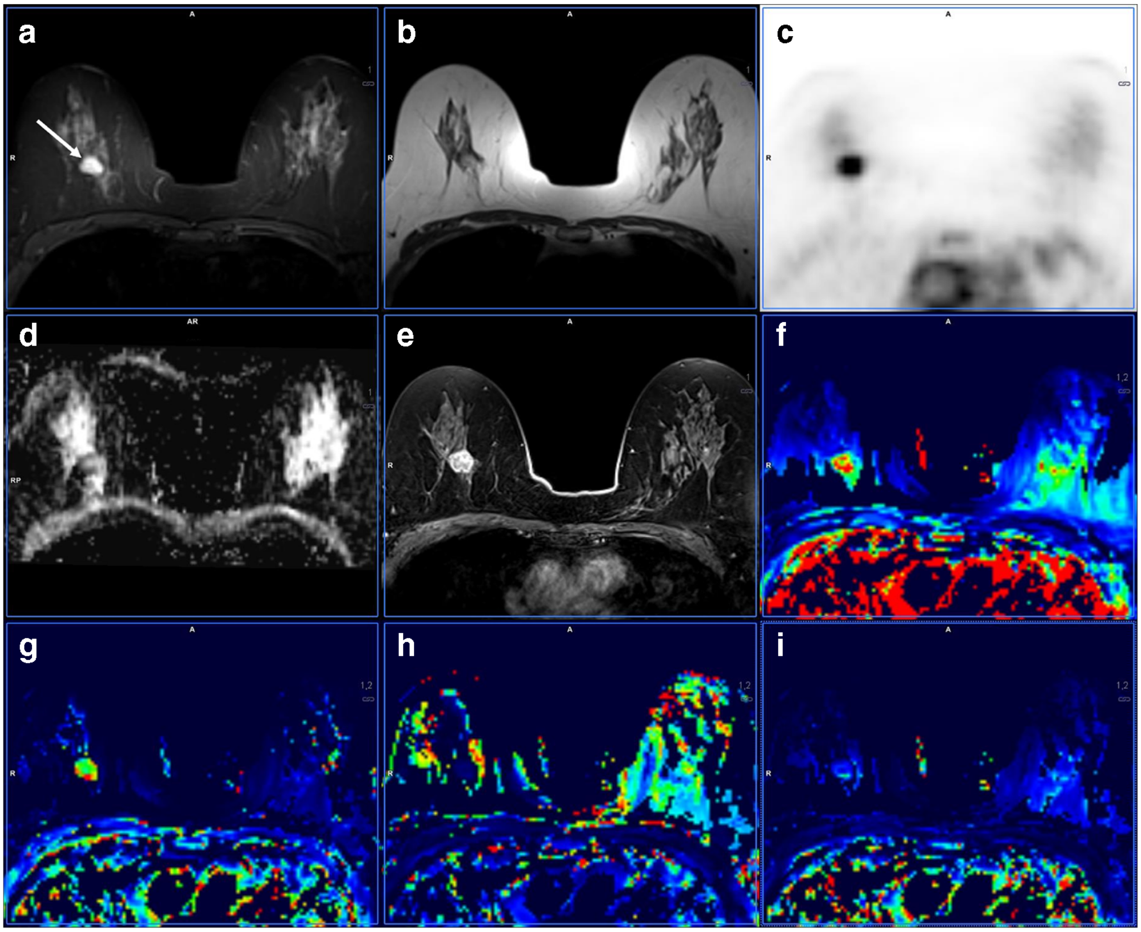

Fig. 1 Hybrid PET/MR breast imaging, axial view with a tumoral lesion (*white arrow*). **a** MR STIR, **b** MR T2 turbo spin echo, **c** FDG-PET, **d** ADC map, **e** T1 3D post-contrast, **f–i** pharmacokinetic map of area under curve (AUC), V_e, K_{ep} e K_{trans}, respectively. Images are from IRCCS SDN

follow-up, in order to know clinical endpoints to stratify the patients based on radiomic differences expressed by the studied patients.

Radiomics is based on the extraction of a large number of characteristics from each lesion of patients in the study. The number of different imaging features that can be obtained is very high, up to several hundreds, often much greater than the number of samples in the study, and this limits the statistical power of the radiomic analysis and potentially generates data over-fitting in the decision models. This issue is already true for a single image modality of hybrid systems and becomes more complex when multimodal studies from hybrid imaging are available. The size of the sample determines the statistical power of the models and should be properly chosen prior to radiomic analyses. Published studies, in particular for PET, evaluated the prognostic or predictive power of radiomics on small-size patient cohorts (< 100), from retrospective data. These exploratory analyses had the great potential to provide the rationale for further investigations in larger cohorts. However, the statistical power of these studies is limited. As a rough indication, it has been estimated that a number of at least 10–15 patients should be included in a radiomic study for each feature when statistical tests are used to test a specific hypothesis [16]. When features are extracted to enter in an AI model, a validation should be provided, and permutation tests could be conducted in case of sample of small size.

Since most of these retrospective studies are based on single-center datasets, in order to avoid false discoveries [32] a validation of the generalizability of the result should be obtained by testing radiomics on other independent datasets. When this is not possible, it is very important to validate the results of such radiomics studies; for example, testing well-known associations among the different groups of patients stratified by radiomics with validated biological markers (e.g., between the stage or subtype and survival [32]), and interpreting the radiomic derived stratifications with respect to such validated results.

In these retrospective studies, it is also very useful to validate the radiomic results by the visual assessment of the different expression of radiomic features on the tumor habitat of single patient images [16]. Textural features can be estimated by calculating a single value for the whole segmented lesion volume or by building a texture map. The VOI-based approach provides a single quantitative descriptor of the texture heterogeneity in the whole structure. The texture map permits the estimation of textural features in each pixel by computing their values considering a square mask of pixels (e.g., 4 × 4) over the image of the lesion. This last method can be useful to clearly show, on the lesion image, a feature that is not uniform within the lesion and to identify sub-regions with high heterogeneity. Moreover, it could be helpful for voxel-based correlations with dose maps in radiotherapy applications [33, 34].

It is, however, clear that radiomics obtain advantages from prospective studies, since such protocols better account for the collection of omics and clinical data in a proper sample size. In these cases, better care could also be taken in the standardization of the imaging and omics studies and in the collection and archive of images, biological data, and clinical information, in order to avoid misleading results due to instrument differences between centers.

Image acquisition and reconstruction

Since the introduction of hybrid imaging, PET/CT has undergone a steady evolution permitting the definition of image acquisition and reconstruction protocols for both the PET and CT components, suitable for oncological, neurological, and cardiological studies respectively. Image quality of studies from current-generation PET7CT systems has achieved a comparable level. However, a large variability is still present in PET/CT images from a quantitative point of view, as a result of the use of different hybrid systems. This is even truer for the PET/MRI hybrid systems, where the technological evolution has been completed for their use in the clinical environment more than for their reproducibility.

Recently, methodological studies have been devoted to the evaluation of the impact of the image acquisition and reconstruction protocols on radiomic features extracted from the medical images, including images from hybrid PET studies. These effects are important in particular in case of multicentric radiomic studies. Some studies have been focused to evaluate the dependence of radiomics features from the physical characteristics of the hybrid PET scanners (e.g. spatial resolution, sensitivity, signal-to-noise ratio), acquisition parameters (e.g. field of view, presence / absence of compensation of the patient physiological movements), or reconstruction algorithms (e.g. number of subsets of iterative methods). These studies have been performed on real images of oncological patients [e.g. 35–38], on simulated images [e.g. 39], on images of phantoms acquired under ideal conditions [e.g. 40] or in conditions miming clinical situations of interest [e.g. 41] and showed that radiomics features are influenced by the image acquisition and reconstruction settings and this influence opens up several issues related to the quality, accuracy, reproducibility, and consistency of the extracted features, and more in general, the results obtained in different studies.

In retrospective multicentric studies when acquisition/reconstruction protocols of patients cannot be harmonized because only reconstructed images have been stored for legal purposes, a methodological study should be performed to guide the selection of those quantitative imaging features more stably with respect for the different acquisition/reconstruction parameters [e.g. 8]. Alternatively, in PET hybrid imaging, some radiomic studies have considered as harmonized images those PET/CT images acquired and reconstructed according to the European Association of Nuclear Medicine (EANM) guidelines [42] and then processed in order to minimize differences between semiquantitative evaluations [43]. Recently, a harmonization method initially proposed for genomic data was optimized and applied to harmonizing radiomic features extracted from the PET images of different hybrid PET systems, with promising results [44]. Although it needs further confirmation through more in-depth studies that also include other imaging modalities, this approach could pave the way for the development of harmonization methods to be applied to retrospective studies that could restore an enviable wealth of data for radiomic analyses.

Image conversion and correction

Data conversion is the transformation of the values in the image data into values of biological meaning. In hybrid PET studies, PET images should be converted into standard uptake value (SUV), a standard semi-quantitative index expressing the measure of the PET radiotracer uptake in the tumor tissues normalized for the radioactive dose injected into the patient [45]. However, proper measurement and annotation of the radioactive injected dose and residual as well as time of injection should be performed for SUV conversion, and this procedure is still not part of routine clinical studies requiring the use of additional time. For this reason, SUV conversion could not be possible for retrospective studies, limiting the application of radiomics.

CT images can be more easily transformed into Hounsfield units (HU), expressing the different tissue densities within the tumors, while in most of the multi-parametric MRI acquisitions no data conversion is applied, since no calibration with functional standard units is usually performed, even if data need to be normalized, as explained in the sections below.

Artifact correction for metal implants should be performed in CT [e.g., 46] while MRI images should be corrected for magnetic field non-uniformity corrections [47].

PET images should be corrected for partial volume effect (PVE), since PVE influences the results of radiomic analyses on PET images. PET PVE occurs when the size of a cancer lesion is comparable with the PET spatial resolution, and is the cause of severe errors in quantitation. Several strategies have been proposed for PET PVE correction; most of them are based on the single PET component [e.g., 48–50]. However, CT and MRI co-registered image can provide a suitable anatomical reference for recovering the actual uptake on the corresponding PET image. MRI-guided PVE corrections have been proven to reduce bias and coefficient of variation of PET images; thus, it could be used to improve robustness of radiomics features. Among the several PVE correction methods using MRI as prior information, the geometric transfer matrix (GTM) method has proven effective [51], where MR images are segmented to different non-overlapping regions representing different tissue types and these regions are then used to correct the PET images. To mitigate the challenge of segmenting the entire object, projection-based tissue activity estimation methods have been proposed which only required the segmentation of a small number of tissues within a small region around the lesion [52–54]. With the same purpose, a filtering method can be used on hybrid PET/MRI systems [55] in which the MR image is used as a prior term on the voxel level.

Image segmentation

Radiomics requires the preliminary identification of a volume of interest (VOI) or sub-volumes of interest within the lesions, representing physiologically distinct volumes (habitats) of potential prognostic value [56].

VOI segmentation is a complex task in medical imaging, and it currently represents a distinct open research topic. A variety of segmentation approaches have been proposed [e.g., 57–63] and this overview does not aim at describing them. A shared opinion regarding segmentation is that each segmentation approach should be chosen after proper evaluation of signal and noise characteristics of the specific image modality. Thus, in hybrid imaging, distinct segmentation methods could be considered for the functional with respect to the structural imaging component. For example, manual contouring from radiologists is the most common praxis accepted in clinical settings since it is a reliable choice for such high-resolution imaging modalities, even if there is as yet no consensus on the best segmentation method for measuring the volume of a lesion on CT and MRI studies. This procedure is more criticized in PET and in functional MRI (e.g., DWI) since it can be affected by a higher level of subjectivity from operators due to the lower resolution and contrast of these images.

Furthermore, the same segmentation approach could be sub-optimal for segment lesions located in different organs even using the same image modality: segmenting a breast lesion from a PET study can have a different accuracy to that obtained when segmenting a lung lesion from PET images.

The segmentation approaches used in published radiomic studies range from manual contouring to semi-automatic up to automatic segmentation methods, consistently with the wide methodology offered by the many years of research on medical image segmentation. Automatic and semi-automatic approaches include thresholding-based methods [59, 60], region growing methods [61], stochastic and learning-based approaches [62], and boundary-based methods [63].

Irrespective of the methods adopted, what has emerged as more important for avoiding bias in radiomic analyses is the good reproducibility and consistency of the segmentation rather than the accuracy. Several studies have examined the impact of segmentation on radiomics [64–67]. These studies showed that the differences in VOI extraction due to different operators or segmentation algorithms cause variations in radiomic features that could generate statistically significant results interpretable of biological meaning [68]. This causes lack of robustness and reproducibility in the whole radiomic process and needs standardization. Since full automation is not always possible, a good acceptable compromise seems to be the use of semi-automatic segmentation methods to be eventually adjusted by expert radiologists, and then the selection of those radiomics features that are found to be more stable with respect to the different expert radiologists. Another criterion to reduce the impact of segmentation in radiomic studies is to apply different segmentation algorithms to obtain the lesion volume, and to select those radiomic features that are more stable with respect to the different segmentation algorithms [69].

However, hybrid-imaging modalities can offer several advantages for the segmentation task coming from the use of the images of different modalities. The VOI can probably be defined on the structural modality (CT or MR) and then propagated for suitability on the PET images, since the inherent co-registration of PET and structural studies can be exploited for a reliable spatial correspondence between modalities [70]. In particular, simultaneously acquired MRI offers several methods for the improvement of PET images [71]. Moreover, long-lasting PET imaging acquisitions may suffer from involuntary movements which occurred in the meantime; therefore, simultaneous fast MRI sequences can also be exploited to estimate the motion and retrospectively correct PET data. Different methods have been proposed in literature to account for both cyclic and non-cyclic motion induced by respiration and gross movements respectively; under these conditions, MRI-driven motion correction has been demonstrated to improve detectability and quantitation of PET-avid lesions [72, 73], with great advantage for propagating anatomical VOI on PET images for the estimation of radiomic features.

Interpolation, re-segmentation, and discretization

In order to obtain features that could be considered rotationally invariant and in order to be able to compare data from different cohorts, images can be resampled for isotropic voxel spacing, and this implies interpolation of intensities in resampled voxels. An operation of rounding on the intensity values may be required after interpolation. This in particular is the case for CT images where signal intensity must be expressed in integer values directly related to tissue attenuation coefficients.

The use of isotropic voxels can be important to guarantee reproducibility across different scanners. Upsampling and downsampling schemes have different advantages or disadvantages, and currently no clear indications exist about which could be the preferable scheme [74–78].

Since interpolation affects image intensity, re-segmentation could be required in particular for intensity-based segmentation masks. In hybrid PET imaging studies, re-segmentation could be required in order to exclude voxels with low signals from the calculation of intensity-based features. In MRI, outlier intensities could be removed by using the Collewet normalization, making it possible to removie from the VOI the voxels outside a specific range of intensities ($[\mu - 3\sigma, \mu + 3\sigma]$, where μ is the mean and σ the standard deviation of grey levels of voxels in the segmented VOI [40].

As a last step before feature extraction, the image intensities within the VOI have to be discretized to practically compute the radiomic features. This operation contributes also to effectively reducing noise.

There are two main schemes for performing discretization. With the first approach, the signal intensities are resampled in a fixed number of bins, while with the second method the signal intensities are resampled in a variable number of equally spaced bins but with a fixed width. Both approaches are commonly used, and some works have made an effort to figure out which approach is preferable in the different image modalities. In hybrid PET imaging, some authors recommend the use of fixed bin size, as a more robust, repeatable method, less sensitive to segmentation and reconstruction changes [75, 77]. Other authors prefer fixed bin number, since this is less correlated with the lesion volume (e.g., 64 bins are recommended since this is sufficient to cover SUV ranges of oncological lesions with 0.25 increments [78, 79].

Some recommendations for discretization algorithms are proposed for each modality in [19] in order to optimize feature inter- and intra-sample reproducibility. Discretization based on fixed bin number is not recommended when the re-segmentation cannot be defined (e.g., on images in arbitrary units).

Quantitative macroscopic image features in hybrid imaging: are they still considered in radiomics?

Among the most popular image measures adopted in the first radiomic studies, we found, were features extracted from a lesion as a whole. Some of these metrics have been now classified within radiomics as "local intensity features" [19], but, with more popular names, have been used and considered for decades as standards for quantitative assessment and reporting in oncology, having being introduced in recent years also in cardiology [e.g., 80] and in rheumatology [e.g., 81, 82].

However, these macroscopic features are not able alone to properly reflect intra-tumor heterogeneity explaining different phenotypes; they are still considered in recent radiomic studies, for several reasons. First of all, clinical recommendation and guidelines still suggest the use of these standard quantitative metrics for the evaluation of response to therapy [e.g., 83–85]. Indeed, at present, there are no results on the level of variations of textural features to be considered for the evaluation of treatment response. In other radiomic studies, the goal is to understand the advantage of textural features compared to standard macroscopic indices with respect to the diagnosis and prognosis of patients, by evaluating their possible complementary and synergistic role [79]. Currently, there are still several radiomic studies that combine macroscopic indices with textural or other radiomic indices [e.g., 60, 86–91]. Another reason for considering macroscopic indexes for lesion characterization is that textural parameters are derived on a series of neighboring voxels, thus imposing limitations on the lesion size to be extracted. This limitation has a particular impact on PET images due to the poor spatial resolution. For PET, in Orlhac et al. [39], a limitation of 5 cc volume was indicated on the basis of a minimum number of neighboring voxels in the *x*, *y*, or *z* directions, and accounting for spatial resolution.

Macroscopic image features in CT studies

In CT studies, linear, cross-sectional, and volumetric measures of tumor size, in particular anatomical tumor volume (ATV), have been for years the most common indices for quantifying cancer aggressiveness and response to treatment [e.g., 92]. Some radiomic studies, in particular in the early days, used ATV together with other advanced features for prediction of response to therapy or recurrence [e.g., 93] or for disease characterization [e.g., 94]. Some studies showed also a role in prognosis and response to treatments of semi-quantitative measures describing the tumor (e.g., space locations, size, shape, lobulation, concavity, irregularity, border definition), the surrounding tissues (vascular convergence, fibrosis periphery) and the tumor-associated findings (nodules, vascular involvement). These descriptors are defined, evaluated, and ranked by independent expert radiologists, often adapted from the Imaging Reporting and Data Systems (I-RADS) of the Colleges of Radiology and from the lexicons of the reference societies. Thanks to the high spatial resolution of CT images, tumor-size features are usually able to be defined manually [e.g., 95].

Macroscopic image features in PET studies

Standardized uptake value (SUV), defined as the mean or the max value of radiotracer uptake in a ROI normalized to patient characteristics, has been the most used quantitative index in PET studies. In addition to SUV metabolic tumor volume (MTV), accounting for the total metabolically active volume of a lesion was considered as a macroscopic PET image feature. SUV peak, that is the mean of SUV in a fixed-size region of interest around the maximum uptake value, is another standard quantitative measurement, and has been recommended for both clinical diagnosis and treatment response [84]. This index was considered more reliable and stable with respect to the maximum value since it is less vulnerable to statistical noise; at the same time, it is less impacted by the technique used for tumor delimitation. Another PET measure, derived from the two mantioned above, is total lesion glycolysis (TLG), which is the product of MTV and mean SUV. All these measures have been proven feasible in many oncological studies; they have been found to be increased in patients with poor prognosis, and decreased in patients responding to effective therapies [96]. However, they suffer from different drawbacks. The most important limitation is related to the partial volume effect (PVE) [e.g., 5, 6] occurring within the size of a cancer lesion, which is comparable with the PET spatial resolution. This effect impairs the accuracy of quantification of MTV and the other-MTV-derived indexes, which can be strongly underestimated for lesions of a few centimeters in diameter. In order to compensate for this effect, some authors have considered only large tumors (e.g., diameter > 2 cm) [e.g., 97]; others have used reconstruction methods incorporating spatial resolution recovery [3]. PVE compensation has proved to increase the statistical significance of correlations among PET MTV, SUV, and TLG and biological prognostic indexes or clinical outcome [98–100]. No studies on the impact of PVE correction on radiomic descriptors are published to our knowledge.

In different studies both using hybrid PET/CT or PET/MRI scanners or by using images PET and CT or MRI from single scanners, PET quantitative macroscopic indexes have been used in combination with CT or MRI macroscopic features in order to evaluate their relationship and their synergic role for diagnosis or for the evaluation of prognosis or response to therapy [e.g., 60, 101–103].

Macroscopic image features in MRI studies

Considering multiparametric MRI, an effort was made to extract quantitative parameters to be used in clinical settings. From T1- and T2-weighted, morphological and tissue characteristics should be explored by quantifying longitudinal and transverse relaxation. Despite the efforts to quantify tissue T1 and T2 values, current clinical cancer diagnosis and monitoring is still based on qualitative visual inspection [104].

T1 contrast was considered quantitatively in dynamic contrast-enhanced- (DCE-) MRI. DCE-MRI studies are T1-weighted acquisitions performed by following the intake of exogenous contrast agents, paermitting imaging of tumor angiogenesis to be performed. Several pharmacokinetic models have been developed to describe perfusion in the tissue microstructure, measuring contrast media passage in tumor tissues and several kinetic parameters such as *K*trans (volume transfer coefficient) and V*e* (extracellular volume ratio) were proposed both calculated at voxel level or as a mean in the tumor. Mean values of such parameters are still evaluated to characterize tumors at the baseline diagnosis as well as response to therapy [e.g., 105–107].

Functional diffusion-weighted MRI (DWI), measuring impeded water diffusion due to tissues microstructural organization, has emerged as a powerful diagnostic tool. The degree of impeded water diffusion can be measured in DWI by the apparent diffusion coefficient (ADC), obtained by fitting of the signal decay observed at different diffusion weightings. Mean value of ADC in tumor has been used for tracking both tumor progression and response to treatment, with promising results in different tumor malignancies [e.g., 60, 108–110], since it expresses tissue cellularity and integrity of cell membranes [111]. With respect to PET, a few methodological works have been devoted to address, examine, and eventually solve technical issues related to the extraction of ADC from DWI-MRI images and to assess ADC quantitative accuracy [112, 113]. Only in recent years have international initiatives been promoted to study from a methodological point of view and to standardize the extraction of such macroscopic measures, as well as to evaluate their accuracy with respect to acquisition and volume definition [114].

In addition to oncology, ADC values are also used in neurology for evaluation and characterization of brain alterations and abnormalities [115–118].

In recent years, the radiomics approach has been widely applied to multiparametric MRI [e.g., 119–121], and high-throughput methods for imaging biomarkers extraction have been applied to the different acquisitions and in different malignancies. Despite this, in many MRI radiomic studies macroscopic metrics are still considered, and macroscopic parameters have been included in prediction models since their use in clinics, due to the lack of studies showing their effectiveness with respect to advanced radiomics indices [e.g. 122–124].

Feature extraction

There are different methods for extracting radiomic features. The initial lack of standardization in terms of nomenclature, definitions of features, and methods for radiomic feature calculation [79] has been resolved, and a radiomic feature standardization is now described in a consensus-based framework on radiomics [19], proposing the classification of radiomic features into a number of families with clear computational methods that include morphological and statistical features and filtering techniques.

In general, image features can be computed according to two main approaches: the first, which is the one widely adopted in the approaches discussed in this manuscript, is based on hand-crafted features, whilst the second employs automatic descriptors generated by deep neural networks.

Hand-crafted features in radiomics leverage on morphological and textural analysis, usually integrated together to improve performance. Morphological features compute measures describing the size and shape of the ROI either in 2D or 3D, which are therefore independent from the gray level intensity distribution. Textural features are calculated from matrices describing the presence of patterns in the images, and they are based on statistical features and on other descriptors capturing texture information. Statistical features compute first- and second-order measures: the former consists of several statistical moments of the first-order image histogram, such as mean, standard deviation, skewness, and kurtosis. The latter are derived from the the gray-level co-occurrence matrix (GLCM), which describes the second-order joint probability function of an image: it is a matrix whose row and column numbers represent gray values, and the cells contain the number of times corresponding grey values are in a certain relationship (angle, distance). GLCM-based features such as second order entropy, energy (also defined as angular second moment) cluster shade, contrast homogeneity, dissimilarity, and correlation describe the heterogeneity and local variation in the image. Furthermore, as mentioned before, a plethora of other textural image features exists, such those now described. A gray level run-length matrix (GLRLM) is a two-dimensional matrix where each element *(i,j)* counts the number of homogeneous runs of *j* voxels with gray intensity *i* in the specified direction. The gray level size zone matrix (GLSZM) is a matrix in which the element at row *i* and column *j* corresponds to the number of homogeneous zones of *j* voxels with the intensity *i*, therefore providing information on the size of homogeneous zones for each gray level. In the neighborhood gray-tone difference matrix (NGTDM), the *i*th entry is a summation of the differences between all pixels with gray-tone *i* and the average value of their neighborhoods specified by a given mask [125]. Different features can be computed from such matrices, such as the coarseness, contrast, entropy, etc. The local binary pattern (LBP) is a grayscale texture descriptor assigning to each pixel of the image a label obtained comparing such a pixel to each of its eight neighbors along a circle, which can be run clockwise or counterclockwise. Different types of LBP operators can be computed varying both the pixel arrangement and the number of neighbors [126]. Readers interested in feature extraction in radiomics can also refer to pyradiomics, i.e., an open-source Python package for the extraction of 2D and 3D radiomics features that aims to establish a reference standard in this field for easy and reproducible feature computation [127].

Filtering techniques transform images by linear or nonlinear filters to reduce noise, and then statistical methods are applied to extract imaging features at different levels. Filters include square, square root, logarithm, exponential filters, wavelet, and Laplacian of Gaussian (LoG).

In several fields, without being limited to radiomics or medical imaging, the rapid uptake of deep learning has improved not only predictive power but also the ability to generate automatically optimized high-level features [128]. While in several cases feature extraction and model fitting take place in a unified step, there is a consensus that the generic descriptors extracted from the deep neural networks are very powerful, as recent results have indicated. Therefore, deep features are used by some authors to feed different traditional machine-learning algorithms [129–131]. We do not describe more deep features, since interested readers can find more details in other manuscripts of this special issue that are specifically directed towards deep learning in radiomics. This choice is also motivated by the fact that that all the papers on decision models in hybrid imaging that we will introduce in the next section make use of hand-crafted features.

A large number of characteristics can be calculated according to the suggested methodology. However, as it is not yet clear what are the most useful features to explain significantly the differences in phenotypes, in any work of radiomics, it is suggested that there should be an explicit description and annotation of which features have been considered by following specific indications and a required nomenclature, including specifications relative to all the steps of the workflow described for the radiomic analysis. For example, it should be reported which features are calculated with reference to feature definitions, or otherwise described in detail, and the specific settings used for feature calculation (e.g., image modality, interpolation, resegmentation, discretization method). This good praxis should facilitate the evaluation and comparison of future findings coming from the research and clinical communities studying radiomics.

The number of features that can be extracted depends on the variety and size of patient data available to train the model.

For this purpose, multicenter studies from multiple hospitals can be more effective than single-center studies. However, collecting data from multiple hospitals is difficult due to ethical, legal, and administrative issues. In order to avoid these problems, it is possible to use a distributed learning approach, i.e., learning from data without the data leaving the hospital. There are several techniques that allow distributed learning. Some of these techniques are aimed at the horizontal distribution of data, that is, each center has the same variables but different patients [132, 133]. Other algorithms are focused on vertically distributed data, i.e., different parts of the same patient data reside in different centers [134, 135]. Bayesian networks adapt to both problems [136–139] and have already been used both for binary and continuous variables [140, 141], allowing effective management of missing data that often occurs with multicenter studies.

Decision models in hybrid imaging: how to build smart and truly multi-parametric decision models for radiomics

Feature selection

The spread of "-omics" strategies has strongly changed the way of thinking about the scientific method [142]. Indeed, managing huge amounts of data eventually collected from heterogeneous sources imposes the replacement of the classical deductive approach with a data-driven inductive approach, so as to generate hypotheses from data.

In this respect, data reduction (also known as feature selection) is a crucial step, also because of the sparsity of significant features in the (big) datasets.

Before beginning the modelling of endpoint of interest from radiomic features, exploratory statistical analysis is useful to find out which features are related to the endpoint of interest, but also to rule out those features that are not correlated and/or are redundant. Relevancy, in this context, is the property of being predictive of the endpoint of interest. Redundancy refers to features that are highly associated with others because they describe the same image properties. Some of these features can be removed as they are not independent predictors of the outcome.

As already underlined, radiomics features should be selected with respect to their stability for variation in acquisition/reconstruction protocol and segmentation methods. If a repeated data set is available, a selection of imaging features based on their reproducibility as obtained by the test–retest may be important [68, 143]. This selection can be made by evaluating, for example, the intraclass correlation coefficient [144, 145].

Redundant features can be identified by the analysis of the covariance matrix of the characteristics extracted for the study sample. Clusters of highly correlated features can be reduced to a single representative image feature, usually chosen as the one presenting the highest dynamic range between the subjects under study [e.g. 146]. The dimensionality of the data can also be reduced by analytic methods such as the analysis of the principal components that combine or transform the original features into a new set of features with low dimensions [147]. Principal component analysis, as well as other linear transforms, determines a new subspace of dimensionality m (either in a linear or a nonlinear way) in the original feature space of dimensionality d ($m \leq d$). However, the new features generated by this process break the interpretability of the original features, a situation that many researchers do not accept. One popular supervised feature selection method is the minimum redundancy maximum relevance (mRMR) method, based on the mutual information between a set of features and the outcome variable. mRMR ranks the input features by maximizing the MI with respect to outcome and minimizing the average MI of higher-ranked features [12].

Nearly all radiomics works explore the data in a supervised fashion and, in this respect, most of the studies performing the textural analysis on MRI and PET analyze statistical correlation, by analyzing if radiomic features are different between groups of patients with different outcomes. According to a widely used taxonomy distinguishing feature selection methods based on their outcomes [148], they make use feature ranking methods, such as appropriate parametric (e.g., *t*-tests) or nonparametric tests (e.g., Mann–Whitney) [149, 150]. The Spearman's rank test can be used to order to identify features that are inter-correlated, in order to remove them from later analysis [151]. These ranking approaches provide outcomes measuring the degrees of dependency of individual features with respect to the target concept. For example, number non-uniformity from NGLDM on PET was the only independent predictor of pathologic response to neoadjuvant chemotherapy for breast cancer ($p = .009$) [149]. The change of features from [18F] FLT-PET and MRI images, acquired before and after treatment, were ranked according for reproducibility in a test/retest evaluation. A scoring function was used to rank the features, which was a combination of the feature variability between test and retest measured using the Bhattacharyya distance, and its sensitivity in capturing response to the treatment as well as for capturing treatment response for renal cell carcinoma [152].

Model assessment

The generalization performance of a learning method relates to its prediction capability on independent test data. Assessment of this performance is very important not only because it helps researchers to select the learning method or model, but also because it provides us a measure of the quality of the chosen model. For these reasons, this issue is discussed now before sub-section 'Modelling', which deals with modelling approaches. This section aims to offer an overview of how to assess a predictive model mostly to empower the readers to judge the quality of published models, whilst interested readers can refer to [153–158] to deepen this topic. In particular, we now introduce performance metrics, and then we will overview techniques to design the experiments to estimate the generalization performance.

Classification tasks in radiomics can be either dichotomies or polychotomies, with a large predominance of the former over the latter; in particular, the literature in hybrid imaging deals with binary tasks, such as classification of recurrence-free survival [74], predicting response to neoadjuvant chemoradiotherapy in esophageal cancer [91], predicting if a lesion will turn out to be metastasis [159, 160], etc..

The confusion matrix contains raw data representing the behavior of a classification system on a set of samples and, for this reason, it is used to evaluate the performances. In this respect, Table 1 shows the confusion matrix in the case of a binary classification task. A first performance score that can be derived is the classification accuracy, defined as $acc = (TP + TN)/(n^{-} + n^{+})$, where $n^{-} = FP + TN$ and $n^{+} = TP + FN$ stand for the number of samples in the negative and positive classes respectively. However, a model cannot be assessed in terms of classification accuracy only, as it is strongly biased to favour the majority class since class distribution is the relationship between the first and the second column of the confusion matrix [161]. Hence, the analysis of the performances in terms of accuracy only would be misleading when the prior class probabilities are different. As an example of this issue, consider the following meaningless situation where in a dataset only 4% of lesions turn out to be a metastasis (associated to the positive class): in this case, if the model labels all test samples as negative it will achieve an accuracy of 96%, but it will fail on all positive cases. Since any performance measure based on values from both columns will be inherently sensitive to class skew, it would be more interesting to dissociate the errors (or the hits) per class. In this respect, four metrics independently estimate the performances in the two classes; following the notation introduce in Table 1 they are:

- Sensitivity, also named as true positive rate or recall, defined as $\mathrm{TP}_{\mathrm{rate}} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$;
- Specificity, also named as true negative rate, defined as $\mathrm{TN}_{\mathrm{rate}} = \mathrm{TN}/(\mathrm{TN} + \mathrm{FP})$;
- False negative rate, defined as $\mathrm{FN}_{\mathrm{rate}} = \mathrm{FN}/(\mathrm{TP} + \mathrm{FN})$;
- False positive rate, defined as $\mathrm{FP}_{\mathrm{rate}} = (\mathrm{TN} + \mathrm{FP})$;

Since $\mathrm{FN}_{\mathrm{rate}} = 1 - \mathrm{TP}_{\mathrm{rate}}$ and $\mathrm{FP}_{\mathrm{rate}} = 1 - \mathrm{TN}_{\mathrm{rate}}$, two independent pairs are sufficient to describe classifier performances since they are independent of prior probabilities and, hence, they are robust when prior class distribution changes over time or over the training and test sets.

Furthermore, the ROC graph is a two-dimensional graph depicting relative trade-offs between $\mathrm{TP}_{\mathrm{rate}}$ plotted on the y axis and $\mathrm{FP}_{\mathrm{rate}}$ plotted on the x axis. Therefore, ROC space corresponds to a square of unitary side, and the point representing the best performance is (0,1), i.e., the uppermost left corner of the square. ROC curve has the attractive property of being insensitive to changes in class distribution. Since the curve is a two-dimensional plot of classifier performance, a common method to compare classifiers consists in reducing ROC performance to a single scalar value by computing the area under the ROC curve, which is referred to as AUC. This value will always be between 0 and 1, but no realistic classifier should have an AUC less than 0.5, which corresponds to random guessing the output class. If the experiment consists in several runs, it is very important to properly average the ROC curve to plot a single graph reporting the average performance. Indeed, vertical averaging, which takes vertical samples of the ROC curves for fixed FP rates and averages the corresponding TP rates, is appropriate only when the $\mathrm{FP}_{\mathrm{rate}}$ can be fixed by the researcher, or when a single-dimensional measure of variation is desired [162]. Furthermore, in this case the $\mathrm{FP}_{\mathrm{rate}}$ is often not under the direct control of the researcher. Hence, it is better to average ROC curves using an independent variable whose value can be controlled directly, such as the threshold on the classifier scores used to compute the curves themselves. This will produce an average ROC curve with confidence bars in the x and y directions. Finally, it is worth noting that the ROC graph measures the ability of a classifier to produce good relative instance scores: this implies that classifier scores should not be compared across models until they have been properly calibrated [154], i.e., they produce scores lying in the same interval.

Table 1 Confusion matrix of a 2-classes problem

	Actual positive	Actual negative
Hypothesise positive	True positive (*TP*)	False positive (*FP*)
Hypothesise negative	False negative (*FN*)	True negative (*TN*)

To assess the performances for regression models, we swap the accuracy or error computation by, for example, the mean squared error (MSE):

$$MSE = \frac{\sum_{i=1}^{l}\left(\hat{y}_i - y_i\right)^2}{l}$$

This quantity measures the average squared difference between the actual value y_i and the predicted value $\hat{y}_i$ is the forecasted value, with l being the number of samples.

All the metrics mentioned so far are used to estimate the predictive performance of a model on unseen data. In this respect, we now overview the holdout method, different techniques within the bootstrap approach and k-fold cross-validation, which can be used to design the experiments.

The holdout method is the simplest model evaluation technique, which takes a labeled dataset and splits it into two parts referred to as training set and test set. Then, we fit a model to the training data and predict the labels of the test set. Since subsampling without replacement alters the statistics of the sample, we can appeal to stratification, i.e., an approach to maintain the original class proportion in resulting subsets. Nevertheless, random subsampling in non-stratified fashion is usually not a big concern when working with relatively large and balanced datasets. Furthermore, when dealing also with hyperparameter tuning, researcher can appeal to a three-way split that divides the dataset into training, validation, and test datasets. This allows us to have a training-validation pair for hyperparameter tuning, independent from the test set. The holdout method may suffer from pessimistically biasing the estimate of the generalization performance, since it uses only a portion of the dataset to train the model. Furthermore, although reducing the size of the test set (i.e., increasing the size of the training set) may decrease this pessimistic bias, the variance of a performance estimates will most likely increase. Since this could be particularly true in case of relatively small datasets, the holdout method is usually considered to be fine when working with large datasets.

The bootstrap approach generates new data from a population by repeated sampling from the original dataset with replacement [155]. Performances are evaluated on so-called out-of-bag samples, which are the unique sets of instances that are not used for model fitting. In practice, for each bootstrap iteration we create a bootstrap data set by sampling with replacement, which is the same size as the original dataset. As a result, some observations may appear more than once in a given bootstrap data set and some not at all. The remaining samples that were not selected for training are used for testing, and the performances are estimated as the average values on test instances. Such an approach may suffer from a pessimistic bias, since bootstrap samples only contain approximately 63.2% of the unique examples from the original dataset [163]. To cope with this issue, Efron introduced the *.632 Estimate* [163], where the estimated model accuracy is computed weighting by 0.632 and by 0.368 the resubstitution accuracy (i.e., the accuracy estimated on the training set) and the out-of-bag sample accuracy respectively. Since this latter approach can provide an optimistic bias in case of overfitting, in [156] the authors proposed the *.632+ Bootstrap* method, where the weights of the resubstitution and out-of-bag accuracies are computed using some a-priori class distribution of the dataset and the proportion of each class example that the classifier predicts in the dataset.

K-fold cross-validation is another approach to estimate the performances. It is based on the idea that each sample in the dataset has the opportunity of being tested. In each round, we split the dataset into k parts: one part is used for model evaluation, and the remaining k−1 parts are merged into a training subset. Furthermore, when there is the need to tune the model hyperparameters, researchers and practitioners can refer to nested cross-validation, where an inner loop nested within the cross-validation loop is used to select the best predicting model on k-fold cross-validation of the training fold. After model selection, the test fold is then used to evaluate the model performance. This procedure offers a workaround for small-dataset situations that shows a low bias in practice where reserving data for independent test sets is not feasible [164]. In general, cross-validation tries to reduce the pessimistic bias of holdout approach by using more training data, in contrast to setting aside a relatively large portion of the dataset as test data, while maintaining not overlapping the test folds, as opposed to repeated holdout iterations. Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation occurring when the number of folds is equal to the number of samples. In this case, for each iteration during LOOCV we fit a model to $n-1$ samples of the dataset and evaluate it on the single, remaining data point. Although computationally expensive, LOOCV can be useful for small datasets. Furthermore, LOOCV estimates are generally associated with a large variance and a small bias. Several works in the literature discuss how the choice of k affects the variance and the bias of the estimate, but there exists no universal (i.e., valid under all distributions) unbiased estimator of the variance of k-fold cross-validation [165]. Hence, there is interest in setting k to find a trade-off between variance and bias in most cases. While interested readers can refer to [153, 157, 166] to delve into this issue, as a rule of thumb 10-fold cross-validation offers the best trade-off between bias and variance, especially when working with small datasets. Moreover, other researchers found that repeating k-fold cross-validation can increase the precision of the estimates while still maintaining a small bias [167].

The growing interest in deep learning has raised the question of model assessment in case of large datasets where the risk of having high variance is quite small. For this reason, to assess the models in deep learning it is common to use the three-way holdout split approach, which is computationally cheap in comparison with k-fold cross-validation. Indeed, the availability of many samples suggests that the way we split the dataset into training, test, and validation should not affect the measured performances.

Finally, let us note that methods used for model assessment are used also in the feature selection stage mentioned in previous subsection. Indeed, feature selection inside the cross-validation loop reduces the bias through overfitting, but it may lead to an excessively pessimistic estimate because fewer training samples are available. Readers interested in this topic may refer to [168].

Modelling

Once selected on the basis of their informative content as described in sub-section 'Feature selection', radiomic features should be combined in integrative models with other biological or clinical data from the collected patients, to generate proper data models to be applied to the study target.

If the study target is disease characterization, the goal is to evaluate a signature of both imaging and biological features in order to characterize the different tumor phenotypes. If the study target is patient outcome or response to treatment, the goal is to generate a signature of radiomic and/or clinical characteristics to classify and stratify a single patient into groups of patients with different clinical outcome, or into groups of responding and not-responding patients.

More advanced data processing methods can be employed for the best combination of multi-modal and multi-parametric biomarkers measured by hybrid imaging, profiting from an accurate spatial and/or temporal correlation of data. There are several techniques that can be used as decision models, and many publications with different purposes with respect to this review present their different characteristics and radiomic applications. Among these methods, logistic regression, multivariate data analysis and classification, multi-kernel learning, machine learning, and deep learning are able to combine a number of different image descriptors measured by PET or CT and MRI within the same classification model [12, 143, 144, 169].

Logistic regression is a popular supervised classifier in radiomics, possibly because of its simplicity and because it is well suited for mixed data types [170]. It is particularly suited to describing radiotherapy outcomes, where the dose–response usually has-a sigmoidal shape [147]. In a study using MR standard imaging procedures combined with O-(2-[18F]fluoroethyl)-L-tyrosine (FET) PET, for the differentiation between local recurrent brain metastasis and radiation injury, a logistic regression model was iteratively selected using the Akaike information criterion (AIC). AIC is a metric to evaluate goodness-of-fit, developed in information theory to provide balance between fitting the regression model and ability of the model to predict on out-of-sample data, minimizing the risk of overfitting. The diagnostic accuracy for the correct differentiation of radiation injury from brain metastasis recurrence was almost 90% [171]. Logistic regression was used also in [91], which investigates how to predict the tumor response to neoadjuvant chemo-radiotherapy in esophageal cancer (EC) patients.

In [172], the authors studied a different multi-modal image-based approach for predicting post-radiotherapy tumor progression in non-small-cell lung cancer (NSCLC). To this end, they analyzed pre-treatment FDG–PET/CT studies of 27 patients, considering local and loco-regional failures. The authors computed a feature set composed of 32 tumor region measures that are based on SUV or HU, on intensity-volume-histogram, and on texture. The statistical analysis was performed using Spearman's correlation and multivariable logistic regression. In particular, the Spearman's rank correlation coefficient makes it possible to correctly rank the response that determines the usefulness in treatment planning selection of an outcome prediction model [173]. The results attained suggested that computing the image features leveraging on a multimodal strategy provides better performance compared to other approaches.

Another approach exploiting PET/CT and MRI was presented in [174], where the authors mined 216 features computed from the PET component of PET/CT and the three MRI sequences (ADC map from DWI, CE-MRI and T2), plus eight other clinical and histopathological parameters and three other treatment parameters. Although not fully hybrid, this is one of the few attempts that truly compute multimodal information. The authors then used Cox regression for multivariate analysis.

An integrated multi-omics model of PET/MRI based radiomics was proposed for prediction of molecular subtype ccA and ccB in patients with primary clear cell renal cell carcinoma (ccRCC). Radiomic features were reduced using sparse partial least squares discriminant analysis (SPLS-DA); then Fisher's linear discriminant analysis (a method that separates patients into two classes using a linear combination of features) was used to obtain the discriminant model in the low dimensional space, in order to derive strongly predictive radiomic signatures [175].

In unsupervised clustering, the machine-learning algorithm has no prior knowledge of the tumor and outcome. The advantage is to identify a cancer subtype in the database of known patients with similar characteristics to the unknown cancer in terms of features. Consensus clustering is a method that provides consensus across multiple runs of a clustering algorithm by subsampling data, evaluating the stability and the best number of clusters [176]. When applied to PET and MRI radiomics for prediction of breast cancer phenotypes and prognosis, unsupervised clustering created three subgroups of features which were associated with tumor grade, overall stage, breast cancer subtypes, and disease recurrence status [177].

A different approach to managing multimodal images could be to fuse the two image sets. Fusion features are related to multimodal image datasets that can be geometrically aligned by registration. Discrete wavelet transform (DWT) can be applied to the registered datasets to combine the spatial and frequency characteristics of multimodality images. In the work of Vallieres [178] PET and MRI were fused into PET/MRI scans. The weight of MRI wavelet sub-bands in the FDG-PET/MRI fusion process, and the presence or not of inversion of MRI intensity values, were studied as radiomic features applicable only to fused images. A set of features extracted from the fused images was selected through a stepwise forward selection, then they were included in a multivariable logistic regression model. In this analysis, it was concluded that fused features could be superior to using separate features extracted from individual images. The same authors in [159] improved the predictive performance of their texture-based radiomic model by optimizing image acquisition parameters before the computation of texture feature. To this end, they considered a data set of PET and MRI of 30 patients, whilst the endpoint consisted in early prediction of lung metastases in soft-tissue sarcomas. To optimize PET and MR image acquisition protocols, the authors performed computerized simulations of image acquisitions with varying parameters. The experiments were conducted applying the bootstrap approach, and the results identify an optimal set of image acquisition parameters improving prediction performance. Indeed, on the one hand, the model based on textures computed from simulated images acquired with a standard clinical set of acquisition parameters attained an average AUC of 0.84 ± 0.01. On the other hand, the use of an optimal set of image acquisition parameters significantly increases model performance ($p = 0.04$), achieving an average AUC of 0.89 ± 0.01. More recently, PET data were combined with CT or MRI to identify a tumor subregion that is expected to contain cells with similar genotypes and phenotypes (i.e., habitats) with distinct combinations of metabolic activity with cellularity or perfusion [160]. In nuclear hybrid imaging, the CT shows a relationship between high or low metabolic regions with blood volume, cellular uptake, and washout, whilst the PET pharmacokinetic analysis makes it possible to calculate the glucose metabolic rate, reflecting the local glucose consumption [179]. An early attempt in this direction collected data of 13 patients affected by lung cancer and, using a traditional descriptive statistic-based approach, identified four tumor regions with different biologic activity [160].

The presence of hypoxia is an important prognostic factor in many types of cancer. A radiomic signature combining contrast-enhanced CT and 18F-FDG PET features was implemented for the presence of high level of hypoxia, defined in terms of its maximum tumor-to-blood uptake ratio > 1.4 in the 18F-FMISO PET for head and neck cancer. The signature had AUC of 0.833 in the validation subset, and used long-run high-gray-level emphasis; the volume with voxels above 70 Hounsfield units from CT, and the maximum (90th percentile) SUV, and the skewness of SUV from PET [180].

One of the factors that prevents clinicians from accepting these machine-learning methods into clinical practice is that they are perceived as “black boxes”, meaning that it is difficult to determine how they arrive at their prediction. Deep neural networks are particularly difficult to understand due to the large number of interacting, non-linear parts [125, 181]. In order to ease interpretability of radiomics for the clinician, the visualization of maps highlighting regions of the tumor that determine the prediction of the deep-learning classifier have been proposed [127].

What is already smart and truly available?

Radiomic software tools

Although most published studies use in-house developed methods, some research groups have implemented software tools dedicated to medical image analytics to be shared with the scientific community. These tools can be useful: 1) to accelerate competences of groups with more recent skills on the topic, 2) to allow results from different research groups to be reproducible and comparable, and 3) to standardize both feature definitions and computation methods impacting on the reliability of radiomic data [19].

All the products considered in this overview were developed by research teams from universities and laboratories (TexRAD [182], MaZda [183–185], LifeX [186], ePAD [187], HeterogeneityCAD [8], PyRadiomics/Radiomics [188], QuantImage [189], Texture Analysis Toolbox [178], and IBEX [190]).

The majority are designed to analyze CT, MRI, and PET; some of them can also analyze other imaging modalities, such as mammography, radiography, or ultrasound.

The five software systems (TexRAD, MaZda, LifeX, ePAD, IBEX) offer the possibility of segmenting the images of lesions, by algorithms ranging from manual to automatic segmentation (e.g., threshold-based algorithms). The three toolkits (HeterogeneityCAD, PyRadiomics/Radiomics, and Texture Analysis Toolbox) are algorithms developed exclusively for features extraction. They can be embedded in more complete software (e.g., 3D Slicer [191]).

All the considered products can extract morphological, first-, second-, and third-order statistical features, except for ePAD. Four of them (TexRAD, MaZda, PyRadiomics/Radiomics, and IBEX) present the possibility of extracting filtering features.

MaZda is implemented with algorithms to reduce feature dimension (e.g., Fisher score). TexRAD can perform discrimination between high- and low-risk patients.

Table 2 presents some of the software, web platforms, and toolkits available free of charge for the extraction of radiomic features, along with some of their main characteristics. Table 2 does not represent an exhaustive list of the products currently available or in the process of being delivered. Furthermore, considering recent and increased interest in the radiomic field, many other dedicated tools are under development and similar tables have been published elsewhere, for instance as reported in [186, 192].

Table 3 shows commercial software programs, which are also becoming increasingly available, due to the interest of many medical device incumbents as well as newcomers such as commercial spin-off of research groups or de-novo start-up companies. However, this table does not aim to comprehensively include all marketed commercial solutions, which become available at a high pace. Such software programs can be divided into:

- research platforms, i.e., tools enabling the discovery of new signatures by linking quantitative imaging biomarkers, clinical and -omics data as input to clinical endpoints as output. These platforms are usually considered to be non-medical devices in that they do not affect the clinical routine, run usually on independent workstations,

Table 2 Open access software programs for radiomics analysis

Software/toolbox	MaZda [183–185]	lifeX [186]	ePAD [187]	HeterogeneityCAD [8]	PyRadiomics/ Radiomics [188]	QuantImage [189]	Texture analysis toolbox [178]	IBEX [190]
Research group	Institute of Electronics, Technical University of Lodz, Poland	IMIV, CEA, Inserm, CNRS, Univ. Paris-Sud, Université Paris Saclay	Rubin Lab.	V.Narayan, J. Jagadeesan	Dana–Farber Cancer Institute, Brigham and Women's Hospital Harvard Medical School, Boston	University of Apllied Science and Arts, Western Switzerland	M. Valliares	The University of Texas MD Anderson Cancer Center, Houston, Texas
Image modalities	CT, MRI, PET	CT, MRI, PET, ultrasound	CT, MRI, radiography	CT, MRI, PET	CT, MRI, PET	CT, PET	CT, MRI, PET	CT, MRI, PET
Segmentation	YES	YES	YES	NO	NO	NO	NO	YES
Segmentation methods	manual, automatic (threshold, flood-filling)	manual, automatic (threshold, snake)	manual	/	/	/	/	manual, automatic (threshold)
Radiomic features (morphology)	YES	YES	NO	YES	YES	YES	NO	YES
Radiomic features (statistical 1° order)	YES	YES	YES	YES	YES	YES	YES	YES
Radiomic features (statistical 2° order)	YES	YES	YES	YES	YES	YES	YES	YES
Radiomic features (statistical 3° order)	YES	YES	NO	YES	YES	YES	YES	YES
Radiomic features (filtering)	YES	NO	NO	NO	YES	NO	NO	YES
Feature selection	YES	NO	NO	NO	NO	NO	NO	NO
Feature selection methods	1. Fisher score 2. classification error and corr. coeff. 3. mutual informat. 4. minimal classification error of 1-nearest neighbor (1-NN)	/	/	/	/	/	/	/
Decision model	NO	NO	NO	NO	NO	NO	NO	NO

Table 3 Commercial software programs for radiomics analysis

Product Name		TexRAD® Research and Lung [193]	QIDS® [194]	Quantib™ ND/Quantib™ Brain [195]	RadiomiX [196]	Radiology Assistant [197]	iBiopsy® [198]	Accipio Ix [199]	Cardio AI, Lung AI, Breast AI, Liver AI [200]	EVIDENS [201]
Company		FEEDBACK Medical, UK	HealthMyne, US	Quantib, the Netherlands	Oncoradiomics, Belgium	Zebra Medical Vision, Israel	Median Technologies, France	MaxQ AI, Israel	Arterys, US	Imagia, France
Image modalities		CT, PET, MRI, mammography	No limitation specified	MRI	CT, MR, PET, Ultrasound, digital pathology	CT	No limitation specified	CT	MRI, CT, mammography	Not specified
Segmentation		YES	NO	YES	NO	YES	NO	Not specified	YES	NO
Clinical grade (disease/district)		YES (oOncology/lung)	YES	YES (neurodegeneration/brain)	YES (oncology/lung, head & neck)	YES (orthopedics/bone, steatosis/liver, emphysema/lung, cardiovascular/heart & brain)	NO	YES	YES (cardiovascular/heart, oncology/breast/lung/liver,	NO
Research		YES	YES	NO	YES	NO	YES	NO	NO	YES
Research	Radiomic features extraction	NO	YES	NO	YES	NO	YES	NO	NO	YES
	Radiomic features (filtering)	YES	Not specified	NO	YES	NO	Not specified	NO	NO	Not specified
	Feature selection	NO	Not specified	NO	YES	NO	Not specified	NO	NO	Not specified
	Data mining tool or machine learning module	YES	Not specified	NO	YES	NO	YES	NO	NO	YES

and are not used to drive clinical decisions. Their main differentiator from open-access software consists of workflow optimization and efficiency improvements, enabling an automatic, end-to-end seamless processing pipeline. TexRad®, QIDS®, RadiomiX, iBiopsy® and EVIDENS offer research capabilities at a different level, ranging from simple features extraction to image filter application and machine-learning modules. In research mode, these software programs are usually open to process any 3D image, DICOM or not, up to 2D digital pathology images.

- clinically validated software programs, to provide decision support systems (DSSs) based on already discovered signatures and thoroughly validated on large independent datasets, also known as clinical-grade DSSs. In order to use these DSSs in clinical practice, regulatory clearance is usually needed, as they fall within the definition of medical devices in many regulatory systems, e.g., class I or II medical device as a function of their intended use (e.g., mere support to decision versus a computer-aided diagnosis/prognosis). DSSs are usually limited to a specific modality, mostly CT, and to a specific disease in a specific body district: these constraints come primarily from the intended use definition to which these DSSs are subjected to be compliantly marketed.

These research or clinical grade DSSs can be embedded into more comprehensive platforms such as picture archive and communication systems (PACS), hospital information systems (HIS), oncology information systems (OIS) or treatment planning systems (TPS), or can be stand-alone. Usually, existing large medical device producers tend to embed DSSs into their research or clinical solutions, while newcomers often offer their solution as a standalone system.

It is not unusual that large medical device players embed open-access or commercial software programs to provide their customers with the possibility of exploring or exploiting radiomic potential: examples are IntelliSPace Discovery (Philips, the Netherlands) which interfaces to Pyradiomics, Advantage Workstation (GE, Buc, France) which interfaces through a plugin to Quantib™ Brain, or Syngo.via Frontier (Siemens, Erlangen, Germany) which interfaces to RadiomiX.

It is also beneficial to mention the platform (www.envoyai.com) which offers the possibility of sharing applications and, once solutions reach product maturity, of commercializing them.

Radiomic biobanks

Radiomic studies can benefit from large datasets including large number of medical images of patients related to clinical and -omics data. Sharing images and data acquired in cancer studies from multiple centers facilitates the collection and storage of clinical cancer datasets to apply the concept of personalized medicine on big data in oncology [202].

With this aim, several initiatives have been started at international level to promote the creation of imaging biobanks in organized databases. A survey among the European Society Radiology (ESR) members published in 2015 [203, 204] describes 27 imaging biobanks, designed for research and clinical purposes. However, in 80% of these the access to images is restricted.

The American National Cancer Institute (NCI) has supported several initiatives, e.g., Quantitative Imaging Network (QIN), Cancer Imaging Program (CIP) and Informatics technology for Cancer Research (ITCR), with new programs funded for Imaging and Digital Pathology in ITCR. One of the most important for radiomics studies is The Cancer Imaging Archive (TCIA) [205], an archive of medical images of cancer (CT, PET, MRI), collected from patients with common cancer diseases (e.g., breast cancer, lung cancer) accessible for public download together with -omics and clinical data when available. At a European level, some European biobanks, UK Biobank [206] and the German National Cohort [207] have launched new projects aimed at the assessment of medical imaging for personalized medicine.

However, despite the efforts to design and optimize imaging biobanks, these are not yet a reality, and further steps must be taken to ensure quality standards in submitted data and in the integration of imaging data with clinical and -omics data.

A new frontier of data sharing gaining increasing momentum consists of distributed machine learning. This approach akes it possible to create virtual biobanks without the need to centralize the data. Also, it allows for a privacy-preserving solution by eliminating the need to transfer private data out of a hospital. In a nutshell, this approach is based on the mathematical proof that, for a selected subset of machine-learning algorithms (e.g., support vector machine — SVM — or neural network — NN), the model built by centralizing the data in only one biobank is equivalent to the model built by calculating a local model at the biobank of each hospital and creating a global model from each local model by using an iterative approach [208]. Prerequisites for such an approach are technical specifications, such as the definition of an ontology at each biobank (e.g., standardized quantitative image biomarkers and clearly defined endpoints), and formal agreements, such as the ownership of the intellectual property generated by this shared approach.

Radiomics standardization

A first guidance for investigations to meet the need of standardized evaluation in the field of radiomics was provided by Lambin et al. [209]. Experts were brought together in a consensus group by Cancer Research UK (CRUK) and the European Organisation for Research and Treatment of

Cancer (EORTC), and asked to answer the needs of imaging biomarkers (IB) with regard to validation and quality assessment. The group delivered recommendations, in particular enhancing the importance of multicenter studies to warrant high quality of IB in clinical studies, and emphasizing the need to standardize IB and monitor IB precision [144]. This need has also been recognized by other international groups, e.g., the Quantitative Imaging Network [210], The European Society of Radiology [203], the EANM, and the RSNA-QIBA Metrology Working Group [42], and has led to the image biomarker standardization initiative (IBSI), an independent international collaboration born with the aim to provide a consensus-based framework for radiomics workflow, definitions, nomenclatures, benchmark data sets and values, and reporting guidelines [19], in order to guide users on a shared methodology to extract radiomic features and to improve reliability of reporting and reproducibility of radiomic studies. This review has been designed in agreement with guidelines and recommendations suggested in these radiomic standardization initiatives.

Discussion and conclusions

Hybrid imaging systems offer enormous amount of complementary, multi-modal, multi-parametric data, co-registered in spatial distribution and time, to explore intratumor heterogeneity with radiomics at both the anatomical and functional macroscopic scale.

Predictive models based on AI can be built on these radiomic multi-parametric image descriptors to personalize the decision-making for patients, opening a new role for hybrid imaging for the prognosis of diseases.

This is an overview of current challenges and available solutions that can help in translating radiomics and AI applications into smart and truly multi-parametric decision models for hybrid PET imaging.

As a first result, this overview of the literature shows that the advantages offered by the multi-modality of PET/CT and PET/MRI systems can help to compensate for the complexities of translating the radiomic approach into hybrid PET imaging.

For example, in hybrid PET/CT studies, radiomic features extracted from the low-dose CT component of PET/CT can be used to better measure heterogeneity in vascularization, necrosis, or cellularity and the different tumor tissue components than is possible with PET alone, thanks to better spatial resolution, spatial sampling, and signal-to-noise ratio. In hybrid PET/MRI studies, the MRI component can better measure heterogeneity of vessel density, perfusion, and other physiological tissue characteristics of tumors.

CT and MRI images from PET/CT and PET/MRI studies can provide a suitable anatomical reference for recovering the actual uptake on the corresponding PET image affected by severe quantitation errors due to partial volume effects.

Furthermore, the inherent co-registration of PET and CT or MRI images offered by hybrid systems can be exploited for a reliable spatial correspondence between modalities, and used to guide the segmentation tasks across multimodal images.

However, until a consensus with regard to the best procedures to be used in the workflow of radiomic analyses is achieved for hybrid imaging (including image protocols for the acquisition, reconstruction, and the different steps for radiomic processing of images), based on the available solutions gathered from the literature, we would like to suggest that prospective radiomic studies from hybrid PET imaging could consider the following strategies:

- to perform PET/CT image acquisitions according to the EANM guidelines [42]. Ad-hoc consensus for radiomic analysis should be obtained from the patients;
- to process images in order to minimize differences between semiquantitative evaluation [43]. A post-reconstruction harmonization method could be applied directly on the extracted radiomic features as in [19];
- to measure and annotate the radioactive injected dose and residual as well as time of injection for SUV conversion;
- to describe and report denoising and quantitation correction applied prior to radiomic extraction (e.g., noise filtering, PVE correction), the type of interpolation used, and the intensity value rounding when applied;
- to segment VOI lesions using automatic or semiautomatic procedures instead of manual segmentation when possible. Use different segmentation algorithms and/or operators to obtain the lesion VOI and then select those radiomic features that are more stable with respect to the different segmentation algorithms and/or different operators. When the segmentation is complex or fails, the VOI can probably be defined on the structural modality (CT or MR) and then suitability propagated on the PET. CT or simultaneously MRI acquisition can be used for movement correction. All the procedures should be described and reported;
- to apply statistical analysis to select stable features with respect to any source of variability due to differences in the acquisition, reconstruction, and segmentation procedures applied to images used for radiomic analysis;
- to describe and report any re-segmentation which has to be performed due to interpolation;
- to extract radiomics features following harmonized nomenclature, definition, and calculations when possible [19], and report on the software tools used, with a clear description of the used functions;
- to reduce features in order to avoid data redundancy and avoid disproportion between available samples and features. All the procedures should be described and reported.

This overview shows also that macroscopic image features are still considered in radiomics studies with hybrid imaging for several reasons, mainly the need to measure the response to treatments through the comparison over time of those markers that have a standardized range of variability (e.g., for the SUV metrics, a cut-off value of 30% has been accepted for associating the changes to actual metabolic variations [84]).

Thus, until a consensus with rgeard to the tolerated variability of radiomic features has been achieved for the evaluation of prognosis or response to therapy, we suggest continuing to measure macroscopic image features and including them among the candidate features for predictive models. We recommend also in this case to report their formula and measurement methods in the relevant publications.

With regard to the decision models based on radiomics, this overview of the literature shows that the pipeline adopted by most of the available approaches investigated so far consists in the use of machine learning and (supervised or unsupervised) feature selection, prior for the model construction. In this last step, we observe that logistic regression is used by most of the papers. Logistic regression is a discriminative method, and it tries to model by just depending on the observed data while learning how to do the classification from the given statistics. As opposed to discriminative approaches, researchers can appeal to generative supervised learning methods that attempt to fit the conditional posterior probability through the Bayes' rule, or any mathematical formalism derived from there. In this respect, there are already concepts that add some parameters (e.g., clinical ones) to a generative neural network [211] to combine the benefits of memorization and generalization. It is worth noting that not many papers have adopted the generative approach, most likely because generative methods need to model both observed and hidden variables. Moreover, they assume some degree of independence among the model features. Focusing again on discriminative models, regression is one of the available choices, well suited for continuous endpoints such as survival in terms of months. Nevertheless, other well-known methods exist, best appropriated for non-continuous endpoints, such as neural networks, decision trees, support vector machines, and so on, which are widely used in radiomics, but whose potential in application to radiomics of hybrid imaging still has to be explored.

Among these, support vector machine (SVM) uses a boundary to separate data points into two categories by maximizing the margin between the two classes. When the two classes are not linearly separable, they can be transformed using a non-linear kernel into a higher-dimensional feature space where they can be separated by a hyperplane.

Random forests are based on decision trees, a popular concept in machine learning. A group of decision trees is trained, and at each branching an attribute is chosen to split among a randomly selected subset of attributes. After a bag of trees is trained, the most frequently predicted class is taken as the final result.

Neural networks are parallel-distributed, interconnected computational models that consist of many nonlinear elements arranged in order to mimic a neuron network [147].

Machine-learning models are prone to overfitting, and careful feature selection and validation must be performed in order to prevent this. Because some machine-learning methods such as SVM are more robust with regard to overfitting than others such as decision trees, careful design of the experiments is necessary.

Cross-validation, sometimes called rotation estimation or out-of-sample testing, is one of various similar model-validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set, being able also to flag problems such as overfitting or selection bias. It is a method used to estimate extra-sample error. The data are split into k equal-sized parts, and k−1 parts are used for training the model. The last part is used for calculation of the prediction error of the trained model. This process is repeated after rotating the training and validation folds so that all parts are used one time for validation. The overall error is calculated by combining the k estimates of prediction error.

In nested cross-validation, an inner loop nested within the cross-validation loop is used to select the best predicting model on k-fold cross-validation on the training fold. After model selection, the test fold is then used to evaluate the model performance. Thus, nested cross-validation trains and evaluates a model using cross-validation [158]. To estimate the model's predictive performance, several authors in the literature perform multiple rounds of cross-validation using different partitions: this procedure, although it increases variability, reduces bias by averaging over the results. Leave-one-out cross-validation is a particular case where one sample is assigned to the testing set, and the others compose either the training or the validation set. It provides a nearly unbiased estimate using only the original data. Furthermore, LOOCV can be used in conjunction with a .632+ bootstrap validation, that, on the contrary, provides a measure with low variance [32, 35]. When the size of the training set is large, researchers can refer to non-exhaustive cross-validation methods, such as the hold-out approach, which do not compute all ways of splitting the original dataset.

Class imbalance learning, a.k.a. class skew, is another issue that can be observed in radiomic approaches. It refers to the problem that the training set has a disproportion among different classes, and it can occur in both binary and multi-class datasets. In binary datasets, which are the majority of cases in hybrid radiomics, it is usually assumed that the positive and negative samples belong to the minority and majority classes respectively. In multi-class datasets, each class contains a certain proportion of the samples. Since traditional learning algorithms are designed to minimize errors over the majority samples, ignoring or paying less attention to instances of the minority classes, this usually results in poor predictive accuracy over the minority ones. For this reason, in the literature there exist many techniques to combat this issue, including internal approaches which tailor an algorithm to imbalanced data, data-selection approaches (also referred to as resampling approaches), cost-sensitive learning, and ensemble learning [172, 212, 213]. Briefly, internal approaches reinforce the learning towards the minority class by changing the optimization function underlying the algorithms. Therefore, they can be defined as algorithmic approaches and from this change a modification in the decision boundary follows. Data-selection approaches resample the original training set until all the classes are approximately equally represented, and they include oversampling, undersampling, and hybrid sampling. Oversampling methods create new minority class samples according to a given algorithm. One of the most commonly used is the synthetic minority over-sampling technique (SMOTE), which creates "synthetic"'data by operating in the feature space [214]. Undersampling approaches discard samples in the majority class, whilst hybrid methods are a combination of the previous two methods. Cost-sensitive learning, also named as instances weighting method, assigns different weights for the training instances belonging to different classes so that the misclassification of the minority class can be highlighted. Ensemble learning provides a framework to incorporate resampling strategy, weighting strategy, or decision threshold adjustment strategy. They combine several base classifiers that outperform every independent one.

Rather than pre-calculating features and training a shallow machine, in deep learning the system learns the features that optimally represent the data for the problem at hand [215] by non-linearly transforming the representation in each level of the network. For example, the first level may represent edges in an image oriented in a particular direction, while the second may detect patterns in the observed edges, and the third could recognize different objects by analyzing pattern groups [216]. Although deep learning, in principle, removes the need for segmentation, one of the major sources of variability of radiomic features, most of the deep-learning studies (e.g., [217]) use segmentation or labelling of the region of interest for better efficiency and accuracy. The application of deep learning in hybrid imaging is still at its beginning, and it has used for automated detection of pulmonary nodules in PET/CT [218].

Some software tools are free or commercially available for radiomic analyses and can be used to accelerate competences on radiomic analysis processes, to perform more reproducible and comparable radiomic studies, and to obtain stable features. Although, currently, only commercial tools have included AI methods to build decision models based on radiomic signatures (in this overview we have reported on some of them, with their main characteristics), given the current state of radiomics, we cannot state that radiomics already "works" in clinical routine, since further studies on standardization and harmonization of radiomics methodology are still under consideration.

Datasets including large number of medical images of patients related to clinical and -omics data can today be shared for radiomic studies from multiple centers. There are several imaging biobanks organized in databases, designed for research and clinical purposes (in this overview we have cited some of them with). Notwithstanding, in most cases the access to images is restricted, such datasets are very useful to test the robustness and generalizability of radiomic studies performed on a single center dataset.

Finally, it is extremely important to follow the international initiatives carried out by the reference communities to stay up to date on the most recent and continuous developments in radiomics, in particular on how data collection and radiomic analysis procedures are harmonized and standardized. The radiomics research field is extremely recent and is developing so fast that it is difficult to be updated through the published literature. These coordinated and shared initiatives by the scientific and clinical communities involved in the medical imaging sector are the only ones that can effectively synthesize scientific evidence and favor an accurate radiomic translation process for the development of predictive models in a personalized medicine perspective.

Compliance with ethical standards

Isabella Castiglioni declares that she has no conflict of interest. Francesca Gallivanone declares that she has no conflict of interest. Paolo Soda declares that he has no conflict of interest. Michele Avanzo declares that he has no conflict of interest. Joseph Stancanello discloses an interest in Oncoradiomics SA. Marco Aiello declares that he has no conflict of interest. Matteo Interlenghi declares that he has no conflict of interest. Marco Salvatore declares that he has no conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Bettinardi V, Mancosu P, Danna M, Giovacchini G, Landoni C, Picchio M, et al. Two dimensional vs three-dimensional imaging in whole body oncologic PET/CT: a discovery–STE phantom and patient study. Q J Nucl Med Mol Imaging. 2007;51(3):214–23.
2. Bailey DL, Pichler BJ, Gückel B, Antoch G, Barthel H, Bhujwalla ZM, et al. Combined PET/MRI: global warming—summary report of the 6th International Workshop on PET/MRI, March 27–29, 2017, Tübingen, Germany. Mol Imaging Biol. 2018;20(1):4–20. https://doi.org/10.1007/s11307-017-1123-5.
3. Rizzo G, Castiglioni I, Russo G, Tana MG, Dell'Acqua F, Gilardi MC, et al. Using deconvolution to improve PET spatial resolution in OSEM iterative reconstruction. Meth Inf Med. 2007;46(3):231–5.
4. Gallivanone F, Canevari C, Gianolli G, Salvatore C, Della Rosa PA, Gilardi MC, et al. A partial volume effect correction tailored for 18F-FDG-PET oncological studies. Biomed Res Int. 2013;2013:780458. https://doi.org/10.1155/2013/780458.

5. Bettinardi V, Castiglioni I, De Bernardi E, Gilardi MC. PET quantification: strategies for partial volume correction. Clin Transl Imaging. 2014;2(3):199–218. https://doi.org/10.1007/s40336-014-0066-y.
6. Soret M, Bacharach SL, Buvat I. Partial-volume effect in PET tumor imaging. J Nucl Med. 2007;48(6):932–45. https://doi.org/10.2967/jnumed.106.035774.
7. Gerlinger M, Rowan AJ, Horswell S, Math M, Larkin J, Endesfelder D, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012;366(10):883–92. https://doi.org/10.1056/NEJMoa1113205 Erratum in: N Engl J Med. 2012 Sep 6;367(10):976.
8. Aerts HJ, Rios-Velazquez E, Leijenaar RT, Parmar C, Grossmann P, Cavalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat Commun. 2014;5:4006. https://doi.org/10.1038/ncomms5006.
9. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning — data mining, inference, and prediction. 2nd ed. New York: Springer; 2008.
10. Segal E, Sirlin CB, Ooi C, Adler AS, Gollub J, Chen X, et al. Decoding global gene expression programs in liver cancer by noninvasive imaging. Nat Biotech. 2007;25(6):675–80.
11. Leger S, Zwanenburg A, Pilz 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.
12. Parmar C, Grossmann P, Bussink J, et al. Machine learning methods for quantitative radiomic biomarkers. Sci Rep. 2015;5: 13087.
13. Nanni L, Brahnam S, Salvatore C, Castiglioni I. Texture descriptors and voxels for the early diagnosis of Alzheimer's disease. Artif Intell Med. 2019;97:19–26.
14. Mazo C, Alegre E, Trujillo M. Classification of cardiovascular tissues using LBP based descriptors and a cascade SVM. Comput Methods Prog Biomed. 2017;147:1–10. https://doi.org/10.1016/j.cmpb.2017.06.003.
15. Koster MJ, Matteson EL, Warrington KJ. Large-vessel giant cell arteritis: diagnosis, monitoring and management. Rheumatology (Oxford). 2018;57(suppl_2):ii32–42. https://doi.org/10.1093/rheumatology/kex424.
16. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures. They are data. Radiology. 2016;278(2):151169. https://doi.org/10.1148/radiol.2015151169.
17. El Naqa I, Grigsby PW, Apte A, Kidd E, Donnelly E, Khullar D, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. Pattern Recogn. 2009;42(6):1162–71.
18. S. Reuzé, A. Schernberg, F. Orlhac, R. Sun, C. Chargari, L. Dercle, E. Deutsch, I. Buvat, and C. Robert. Radiomics in nuclear medicine applied to radiation therapy: methods, pitfalls, and challenges. Int J Rad Oncology. 2018;102(4):1117–42.
19. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. 2018 [Current version v9 2019]. https://arxiv.org/pdf/1612.07003
20. Sakurai H, Asamura H, Miyaoka E, Yoshino I, Fujii Y, Nakanishi Y, et al. Differences in the prognosis of resected lung adenocarcinoma according to the histological subtype: a retrospective analysis of Japanese lung cancer registry data. Eur J Cardiothorac Surg. 2015;45:100–7. https://doi.org/10.1093/ejcts/ezt284.
21. Mann RM, Kuhl CK, Kinkel K, Boetes C. Breast MRI: guidelines from the European Society of Breast Imaging. Eur Radiol. 2008;18(7):1307–18. https://doi.org/10.1007/s00330-008-0863-7.
22. Kunkel M, Reichert TE, Benz P, Lehr HA, Jeong JH, Wieand S, et al. Overexpression of Glut-1 and increased glucose metabolism in tumors are associated with a poor prognosis in patients with oral squamous cell carcinoma. Cancer. 2003;97(4):1015–24.
23. Basu S, Kwee TC, Gatenby R, Saboury B, Torigian DA, Alavi A. Evolving role of molecular imaging with PET in detecting and characterizing heterogeneity of cancer tissue at the primary and metastatic sites, a plausible explanation for failed attempts to cure malignant disorders. Eur J Nucl Med Mol Imaging. 2011;38:987–91. https://doi.org/10.1007/s00259-011-1863-4.
24. Tixier F, Groves AM, Goh V, Hatt M, Ingrand P, Le Rest CC, et al. Correlation of intra-tumor 18F-FDG uptake heterogeneity indices with perfusion CT derived parameters in colorectal cancer. PLoS One. 2014;9(6):e99567. https://doi.org/10.1371/journal.pone.0099567.
25. Rockwell S, Dobrucki IT, Kim EY, Marrison ST, Vu VT. Hypoxia and radiation therapy: past history, ongoing research, and future promise. Curr Mol Med. 2009;9(4):442–58.
26. Hatt M, Hanzouli H, Rest CCL, Visvikis D. Comparison of edge-preserving filters for unbiased quantification in 18F-FDG PET imaging. J Nucl Med. 2015;56:1828.
27. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. Eur Radiol Exp. 2018;2(1):36. https://doi.org/10.1186/s41747-018-0068-z.
28. Win T, Miles KA, Janes SM, Ganeshan B, Shastry M, Endozo R, et al. Tumor heterogeneity and permeability as measured on the CT component of PET/CT predict survival in patients with nonsmall cell lung cancer. Clin Cancer Res. 2013;19:3591–9. https://doi.org/10.1158/1078-0432.CCR-12-1307.
29. Yoon SH, Park CM, Park SJ, Yoon J-H, Hahn S, Goo JM. Tumor heterogeneity in lung cancer: assessment with dynamic contrast-enhanced MR imaging. Radiology. 2016;280(3):940–8. https://doi.org/10.1148/radiol.2016151367.
30. Michallek F, Dewey M. Fractal analysis in radiological and nuclear medicine perfusion imaging: a systematic review. Eur Radiol. 2014;24:60–9. https://doi.org/10.1007/s00330-013-2977-9.
31. O'Sullivan F, Roy S, Eary J. A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data. Biostatistics. 2003;4:433–48.
32. 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(5):e0124165. https://doi.org/10.1371/journal.pone.0124165.
33. Maani R, Yang YH, Kaira S. Voxel-based texture analysis of the brain. Plos One. 2015;10(3):e0117759. https://doi.org/10.1371/journal.pone.0117759.
34. Scalco E, Rizzo G. Texture analysis of medical images for radiotgerapy applications. Br J Radiol. 2017;90(1070): 20160642. https://doi.org/10.1259/bjr.20160642.
35. Fave X, Cook M, Frederick A, Zhang L, Yang J, Fried D, et al. Preliminary investigation into sources of uncertainty in quantitative imaging features. Comput Med Imaging Graph. 2015;44:54–61. https://doi.org/10.1016/j.compmedimag.2015.04.006.
36. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring computed tomography scanner variability of radiomics features. Investig Radiol. 2015;50(11):757–65. https://doi.org/10.1097/RLI.0000000000000180.
37. Orlhac F, Soussan M, Maisonobe JA, 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(3):414–22. https://doi.org/10.2967/jnumed.113.129858.
38. Becker AS, Wagner MW, Wurnig MC, Boss A. Diffusion-weighted imaging of the abdomen: Impact of b-values on texture analysis features. NMR Biomed. 2017; 30(1). https://doi.org/10.1002/nbm.3669.
39. Orlhac F, Nioche C, Soussan M, Buvat I. Understanding changes in tumor texture indices in PET: a comparison between visual

assessment and index values in simulated and patient data. J Nucl Med. 2017;58(3):387–92. https://doi.org/10.2967/jnumed.116.181859.

40. Collewet G, Strzelecki M, Mariette F. Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magn Reson Imaging. 2004;22(1):81–91. https://doi.org/10.1016/j.mri.2003.09.001.

41. Gallivanone F, Carne I, Interlenghi M, D'Ambrosio D, Baldi M, Fantinato D, et al. A method for manufacturing oncological phantoms for the quantification of 18F-FDG PET and DW-MRI studies. Contrast Media Mol Imaging. 2017;2017:3461684. https://doi.org/10.1155/2017/3461684.

42. Boellaard R, Delgado-Bolton R, Oyen WJ, 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(2):328–54. https://doi.org/10.1007/s00259-014-2961-x.

43. Sollini M, Cozzi L, Antunovic L, Chiti A, Kirienko M. PET Radiomics in NSCLC: state of the art and a proposal for harmonization of methodology. Sci Rep. 2017;7(1):358. https://doi.org/10.1038/s41598-017-00426-y.

44. 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(8):1321–8. https://doi.org/10.2967/jnumed.117.199935.

45. Thie JA. Understanding the standardized uptake value, its methods, and implications for usage. J Nucl Med. 2004;45(9):1431–4.

46. Zhou W, Bartlett DJ, Diehn FE, Glazebrook KN, Kotsenas AL, Carter RE, Fletcher JG, McCollough CH, Leng S. Reduction of metal artifacts and improvement in dose efficiency using photon-counting detector computed tomography and tin filtration. Invest Radiol. 2018;54(4):204–11.

47. Vovk U, Pernus F, Likar B. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans Med Imaging. 2007;26(3):405–21.

48. Gallivanone F, Stefano A, Grosso E, Canevari C, Gianolli L, Messa C, et al. PVE correction in PET-CT whole-body oncological studies from PVE-affected images. IEEE Trans Nucl Sci. 2011;58(3):736–47.

49. Strul D, Bendriem B. Robustness of anatomically guided pixel-by-pixel algorithms for partial volume effect correction in positron emission tomography. J Cereb Blood Flow Metab. 1999;19(5):547–59.

50. Erlandsson K, Buvat I, Pretorius PH, Thomas BA, Hutton BF. A review of partial volume correction techniques for emission tomography and their applications in neurology, cardiology and oncology. Phys Med Biol. 2012;57(21):R119–59. https://doi.org/10.1088/0031-9155/57/21/R119.

51. Rousset O, Rahmim A, Alavi A, Zaidi H. Partial volume correction strategies in PET. PET Clin. 2007; 2(2):235–249. https://doi.org/10.1016/j.cpet.2007.10.005.

52. Zaidi H, Ruest T, Schoenahl F, Montandon ML. Comparative assessment of statistical brain MR image segmentation algorithms and their impact on partial volume correction in PET. Neuroimage. 2006;32(4):1591–607.

53. Moore SC, Southekal S, Park MA, McQuaid SJ, Kijewski MF, Müller SP. Improved regional activity quantitation in nuclear medicine using a new approach to correct for tissue partial volume and spillover effects. IEEE Trans Med Imaging. 2012;31(2):405–16. https://doi.org/10.1109/TMI.2011.2169981.

54. Southekal S, McQuaid SJ, Kijewski MF, Moore SC. Evaluation of a method for projection-based tissue-activity estimation within small volumes of interest. Phys Med Biol. 2012;57(3):685–701. https://doi.org/10.1088/0031-9155/57/3/685.

55. Yan J, Lim JC, Townsend DW. MRI-guided brain PET image filtering and partial volume correction. Phys Med Biol. 2015; 60(3):961–976. https://doi.org/10.1088/0031-9155/60/3/961.

56. Incoronato M, Aiello M, Infante T, Cavaliere C, Grimaldi AM, Mirabelli P, et al. Radiogenomic analysis of oncological data: a technical survey. Int J Mol Sci. 2017;18(4):e805. https://doi.org/10.3390/ijms18040805.

57. Foster B, Bagci U, Mansoor A, Xu Z, Mollura DJ. A review on segmentation of positron emission tomography images. Comput Biol Med. 2014;50:76–96. https://doi.org/10.1016/j.compbiomed.2014.04.014.

58. Zaidi H, Alavi A. Naqa. Novel quantitative PET techniques for clinical decision support in oncology. Semin Nucl Med. 2018;48(6):548–64. https://doi.org/10.1053/j.semnuclmed.2018.07.003.

59. Gallivanone F, Interlenghi M, Canervari C, Castiglioni I. A fully automatic, threshold-based segmentation method for the estimation of the metabolic tumor volume from PET images: validation on 3D printed anthropomorphic oncological lesions. J Instrum. 2016;11(1):C01022.

60. Gallivanone F, Panzeri MM, Canevari C, Losio C, Gianolli L, De Cobelli F, et al. Biomarkers from in vivo molecular imaging of breast cancer: pretreatment 18F-FDG PET predicts patient prognosis, and pretreatment DWI-MR predicts response to neoadjuvant chemotherapy. MAGMA. 2017;30(4):359–73. https://doi.org/10.1007/s10334-017-0610-7.

61. Veeraraghavan H, Dashevsky BZ, Onishi N, Sadinski M, Morris E, Deasy JO. Sutton appearance constrained semi-automatic segmentation from DCE-MRI is reproducible and feasible for breast cancer radiomics: a feasibility study. Sci Rep. 2018;8(1):4838. https://doi.org/10.1038/s41598-018-22980-9.

62. Huang X, Sun W, Tseng TB, Li C, Qian W. Fast and fully-automated detection and segmentation of pulmonary nodules in thoracic CT scans using deep convolutional neural networks. Comput Med Imaging Graph. 2019;74:25–36. https://doi.org/10.1016/j.compmedimag.2019.02.003.

63. Ma Z, Wu X, Song Q, Luo Y, Wang Y, Zhou J. Automated nasopharyngeal carcinoma segmentation in magnetic resonance images by combination of convolutional neural networks and graph cut. Exp Ther Med. 2018;16(3):2511–21. https://doi.org/10.3892/etm.2018.6478.

64. Haga A, Takahashi W, Aoki S, Nawa K, Yamashita H, Abe O, et al. Classification of early stage non-small cell lung cancers on computed tomographic images into histological types using radiomic features: interobserver delineation variability analysis. Radiol Phys Technol. 2018;11(1):27–35. https://doi.org/10.1007/s12194-017-0433-2.

65. Hatt M, Laurent B, Fayad H, Jaouen V, Visvikis D, Le Rest CC. Tumour functional sphericity from PET images: prognostic value in NSCLC and impact of delineation method. Eur J Nucl Med Mol Imaging. 2018;45(4):630–41. https://doi.org/10.1007/s00259-017-3865-3.

66. Huang Q, Lu L, Dercle L, Lichtenstein P, Li Y, Yin Q, et al. Interobserver variability in tumor contouring affects the use of radiomics to predict mutational status. J Med Imaging (Bellingham). 2018;5(1):011005. https://doi.org/10.1117/1.JMI.5.1.011005.

67. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. Int J Radiat Oncol Biol Phys. 2018;102(4):1143–58. https://doi.org/10.1016/j.ijrobp.2018.05.053.

68. 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.

69. Wu W, Parmar C, Grossmann P, Quackenbush J, Lambin P, Bussink J, et al. Exploratory study to identify radiomics classifiers for lung cancer histology. Front Oncol. 2016;6:71. https://doi.org/10.3389/fonc.2016.00071.

70. Monti S, Cavaliere C, Covello M, Nicolai E, Salvatore M, Aiello M. An evaluation of the benefits of simultaneous acquisition on PET/MR coregistration in head/neck imaging. J Healthc Eng. 2017;2017:2634389. https://doi.org/10.1155/2017/2634389.
71. Lalush DS. Magnetic resonance–derived improvements in PET imaging. Magn Reson Imaging Clin N Am. 2017;25(2):257–72. https://doi.org/10.1016/j.mric.2016.12.002.
72. Manber R, Thielemans K, Hutton BF, Wan S, Fraioli F, Barnes A, et al. Clinical impact of respiratory motion correction in simultaneous PET/MR, using a joint PET/MR predictive motion model. J Nucl Med. 2018;59(9):1467–73. https://doi.org/10.2967/jnumed.117.191460.
73. Fürst S, Grimm R, Hong I, Souvatzoglou M, Casey ME, Schwaiger M, et al. Motion correction strategies for integrated PET/MR. J Nucl Med. 2015;56(2):261–9. https://doi.org/10.2967/jnumed.114.146787.
74. Huang SY, Franc BL, Harnish RJ, Liu G, Mitra D, Copeland TP, et al. Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis. NPJ Breast Cancer. 2018;4:24. https://doi.org/10.1038/s41523-018-0078-2.
75. Leijenaar RT, Nalbantov G, Carvalho S, van Elmpt WJ, Troost EG, 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. https://doi.org/10.1038/srep11075.
76. 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(12):e0145063. https://doi.org/10.1371/journal.pone.0145063.
77. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing agreement between radiomic features computed for multiple CT imaging settings. PLoS One. 2016;11(12):e0166550. https://doi.org/10.1371/journal.pone.0166550.
78. 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. https://doi.org/10.2967/jnumed.111.099127.
79. Buvat I, Orlhac F, Soussan M. Tumor texture analysis in PET: where do we stand? J Nucl Med. 2015;56(11):1642–4.
80. Besenyi Z, Ágoston G, Hemelein R, Annamária B, Nagy FT, Varga A, et al. Detection of myocardial inflammation by 18F-FDG-PET/CT in patients with systemic sclerosis without cardiac symptoms: a pilot study. Clin Exp Rheumatol. 2018.
81. Berti A, Della-Torre E, Gallivanone F, Canevari C, Milani R, Lanzillotta M, et al. Quantitative measurement of 18F-FDG PET/CT uptake reflects the expansion of circulating plasmablasts in IgG4-related disease. Rheumatology (Oxford). 2017;56(12):2084–92. https://doi.org/10.1093/rheumatology/kex234.
82. Mabey E, Rutherford A, Galloway J. Differentiating disease flare from infection: a common problem in rheumatology. Do 18F-FDG PET/CT scans and novel biomarkers hold the answer? Curr Rheumatol Rep. 2018;20(11):70. https://doi.org/10.1007/s11926-018-0779-4.
83. Schwartz LH, Seymour L, Litière S, Ford R, Gwyther S, Mandrekar S, et al. RECIST 1.1 — standardisation and disease-specific adaptations: perspectives from the RECIST Working Group. Eur J Cancer. 2016;62:138–45. https://doi.org/10.1016/j.ejca.2016.03.082.
84. Wahl RL, Jacene H, Kasamon Y, Lodge MA. From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. J Nucl Med. 2009;50(Suppl):122S–50S. https://doi.org/10.2967/jnumed.108.057307.
85. Vargas HA, Hötker AM, Goldman DA, Moskowitz CS, Gondo T, Matsumoto K, et al. Updated prostate imaging reporting and data system (PIRADS v2) recommendations for the detection of clinically significant prostate cancer using multiparametric MRI: critical evaluation using whole-mount pathology as standard of reference. Eur Radiol. 2016;26(6):1606–12. https://doi.org/10.1007/s00330-015-4015-6.
86. Carvalho S, Leijenaar RTH, Troost EGC, van Timmeren JE, Oberije C, van Elmpt W, et al. 18F-fluorodeoxyglucose positron-emission tomography (FDG-PET) — radiomics of metastatic lymph nodes and primary tumor in non-small cell lung cancer (NSCLC) — a prospective externally validated study. PLoS One. 2018;13(3):e0192859. https://doi.org/10.1371/journal.pone.0192859.
87. Li K, Sun H, Lu Z, Xin J, Zhang L, Guo Y, et al. Value of [18F]FDG PET radiomic features and VEGF expression in predicting pelvic lymphatic metastasis and their potential relationship in early-stage cervical squamous cell carcinoma. Eur J Radiol. 2018;106:160–6. https://doi.org/10.1016/j.ejrad.2018.07.024.
88. De Bernardi E, Buda A, Guerra L, Vicini D, Elisei F, Landoni C, et al. Radiomics of the primary tumour as a tool to improve 18F-FDG-PET sensitivity in detecting nodal metastases in endometrial cancer. EJNMMI Res. 2018;8(1):86. https://doi.org/10.1186/s13550-018-0441-1.
89. van Helden EJ, Vacher YJL, van Wieringen WN, van Velden FHP, Verheul HMW, Hoekstra OS, et al. Radiomics analysis of pre-treatment [18F]FDG PET/CT for patients with metastatic colorectal cancer undergoing palliative systemic treatment. Eur J Nucl Med Mol Imaging. 2018;45(13):2307–17. https://doi.org/10.1007/s00259-018-4100-6.
90. Parvez A, Tau N, Hussey D, Maganti M, Metser U. 18F-FDG PET/CT metabolic tumor parameters and radiomics features in aggressive non-Hodgkin's lymphoma as predictors of treatment outcome and survival. Ann Nucl Med. 2018. https://doi.org/10.1007/s12149-018-1260-1.
91. Beukinga RJ, Hulshoff JB, Mul VEM, Noordzij W, Kats-Ugurlu G, Slart RHJA, et al. Prediction of response to neoadjuvant chemotherapy and radiation therapy with baseline and restaging 18F-FDG PET imaging biomarkers in patients with esophageal cancer. Radiology. 2018;287(3):983–92. https://doi.org/10.1148/radiol.2018172229.
92. Sohaib SA, Turner B, Hanson JA, Farquharson M, Oliver RT, Reznek RH. CT assessment of tumour response to treatment: comparison of linear, cross-sectional and volumetric measures of tumour size. Br J Radiol. 2000;73(875):1178–84. https://doi.org/10.1259/bjr.73.875.11144795.
93. Coroller TP, Grossmann P, Hou Y, Rios Velazquez E, Leijenaar RT, Hermann G, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol. 2015;114(3):345–50. https://doi.org/10.1016/j.radonc.2015.02.015.
94. Sato Y, Yanagawa M, Hata A, Enchi Y, Kikuchi N, Honda O, et al. Volumetric analysis of the thymic epithelial tumors: correlation of tumor volume with the WHO classification and Masaoka staging. J Thorac Dis. 2018;10(10):5822–32. https://doi.org/10.21037/jtd.2018.09.133.
95. Wang H, Schabath MB, Liu Y, Berglund AE, Bloom GC, Kim J, et al. Semiquantitative computed tomography characteristics for lung adenocarcinoma and their association with lung cancer survival. Clin Lung Cancer. 2015;16(6):e141–63. https://doi.org/10.1016/j.cllc.2015.05.007.
96. Groheux D, Giacchetti S, Espié M, Rubello D, Moretti JL, Hindié E. Early monitoring of response to neoadjuvant chemotherapy in breast cancer with 18F-FDG PET/CT: defining a clinical aim. Eur J Nucl Med Mol Imaging. 2011;38(3):419–25. https://doi.org/10.1007/s00259-010-1660-5.
97. Groheux D, Giacchetti S, Moretti JL, Porcher R, Espié M, Lehmann-Che J, et al. Correlation of high 18F-FDG uptake to clinical, pathological and biological prognostic factors in breast

cancer. Eur J Nucl Med Mol Imaging. 2011;38(3):426–35. https://doi.org/10.1007/s00259-010-1640-9.

98. Gallivanone F, Canevari C, Sassi I, Zuber V, Marassi A, Gianolli L, et al. Partial volume corrected 18F-FDG PET mean standardized uptake value correlates with prognostic factors in breast cancer. Q J Nucl Med Mol Imaging. 2014;58(4):424–39.

99. Giganti F, De Cobelli F, Canevari C, Orsenigo E, Gallivanone F, Esposito A, et al. Response to chemotherapy in gastric adenocarcinoma with diffusion-weighted MRI and (18) F-FDG-PET/CT: correlation of apparent diffusion coefficient and partial volume corrected standardized uptake value with histological tumor regression grade. J Magn Reson Imaging. 2014;40(5):1147–57. https://doi.org/10.1002/jmri.24464.

100. Picchio M, Kirienko M, Mapelli P, Dell'Oca I, Villa E, Gallivanone F, et al. Predictive value of pre-therapy (18)F-FDG PET/CT for the outcome of (18)F-FDG PET-guided radiotherapy in patients with head and neck cancer. Eur J Nucl Med Mol Imaging. 2014;41(1):21–31. https://doi.org/10.1007/s00259-013-2528-2.

101. Inglese M, Cavaliere C, Monti S, Forte E, Incoronato M, Nicolai E, et al. A multi-parametric PET/MRI study of breast cancer: evaluation of DCE-MRI pharmacokinetic models and correlation with diffusion and functional parameters. NMR Biomed. 2019; 32(1):e4026. https://doi.org/10.1002/nbm.4026.

102. Vallières M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med Biol. 2015;60(14):5471–96. https://doi.org/10.1088/0031-9155/60/14/5471.

103. Lv W, Yuan Q, Wang Q, Ma J, Feng Q, Chen W, Rahmim A, Lu L. Radiomics analysis of PET and CT components of PET/CT imaging integrated with clinical parameters: application to prognosis for nasopharyngeal carcinoma. Mol Imaging Biol. 2019. https://doi.org/10.1007/s11307-018-01304-3.

104. Lin G, Keshari KR, Park JM. Cancer metabolism and tumor heterogeneity: imaging perspectives using MR imaging and spectroscopy. Contrast Media Mol Imaging. 2017;2017:6053879. https://doi.org/10.1155/2017/6053879.

105. Crowe W, Wang L, Zhang Z, Varagic J, Bourland JD, Chan MD, et al. MRI evaluation of the effects of whole brain radiotherapy on breast cancer brain metastasis. Int J Radiat Biol. 2018;30:1–27. https://doi.org/10.1080/09553002.2019.

106. Syed AK, Woodall R, Whisenant JG, Yankeelov TE, Sorace AG. Characterizing trastuzumab-induced alterations in Intratumoral heterogeneity with quantitative imaging and immunohistochemistry in HER2+ breast cancer. Neoplasia. 2018;21(1):17–29. https://doi.org/10.1016/j.neo.2018.10.008.

107. Sun NN, Liu C, Ge XL, Wang J. Dynamic contrast-enhanced MRI for advanced esophageal cancer response assessment after concurrent chemoradiotherapy. Diagn Interv Radiol. 2018;24(4):195–202. https://doi.org/10.5152/dir.2018.17369.

108. Salvaggio G, Calamia M, Purpura P, Bartolotta TV, Picone D, Dispensa N, et al. Role of apparent diffusion coefficient values in prostate diseases characterization on diffusion-weighted magnetic resonance imaging. Minerva Urol Nefrol. 2018;71(2):154–60. https://doi.org/10.23736/S0393-2249.18.03065-5.

109. Oda T, Sue M, Sasaki Y, Ogura I. Diffusion-weighted magnetic resonance imaging in oral and maxillofacial lesions: preliminary study on diagnostic ability of apparent diffusion coefficient maps. Oral Radiol. 2018;34(3):224–8. https://doi.org/10.1007/s11282-017-0303-y.

110. Li QW, Qiu B, Wang B, Wang DL, Yin SH, Yang H, et al. Prediction of pathologic responders to neoadjuvant chemoradiotherapy by diffusion-weighted magnetic resonance imaging in locally advanced esophageal squamous cell carcinoma: a prospective study. Dis Esophagus. 2018; 31(2). https://doi.org/10.1093/dote/dox121.

111. Hamstra DA, Rehemtulla A, Ross BD. Diffusion magnetic resonance imaging: a biomarker for treatment response in oncology. J Clin Oncol. 2007;25(26):4104–9. https://doi.org/10.1200/JCO.2007.11.9610.

112. Arponen O, Sudah M, Masarwah A, Taina M, Rautiainen S, Könönen M, et al. Diffusion-weighted imaging in 3.0 tesla breast MRI: diagnostic performance and tumor characterization using small subregions vs. whole tumor regions of interest. PLoS One. 2015;10(10):e0138702. https://doi.org/10.1371/journal.pone.0141833.

113. Merhemic Z, Imsirovic B, Bilalovic N, Stojanov D, Boban J, Thurnher MM. Apparent diffusion coefficient reproducibility in brain tumors measured on 1.5 and 3 T clinical scanners: a pilot study. Eur J Radiol. 2018;108:249–53. https://doi.org/10.1016/j.ejrad.2018.10.010.

114. Shukla-Dave A, Obuchowski NA, Chenevert TL, Jambawalikar S, Schwartz LH, Malyarenko D, et al. Quantitative imaging biomarkers alliance (QIBA) recommendations for improved precision of DWI and DCE-MRI derived biomarkers in multicenter oncology trials. J Magn Reson Imaging. 2018;49(7):e101-e121. https://doi.org/10.1002/jmri.2651.

115. Rana L, Sood D, Chauhan R, Shukla R, Gurnal P, Nautiyal H, et al. MR imaging of hypoxic ischemic encephalopathy - distribution patterns and ADC value correlations. Eur J Radiol Open. 2018;5:215–20. https://doi.org/10.1016/j.ejro.2018.08.001.

116. Zhu MJ, Wang Y, Li HJ, Yang M, Mo XM, Cheng R, et al. Brain alteration in neonates with congenital heart disease using apparent diffusion coefficient histograms. Zhonghua Yi Xue Za Zhi. 2018;98(39):3162–5. https://doi.org/10.3760/cma.j.issn.0376-2491.2018.39.007.

117. Seyithanoğlu MH, Abdallah A, Dündar TT, Kitiş S, Aralaşmak A, Gündağ Papaker M, et al. Investigation of brain impairment using diffusion-weighted and diffusion tensor magnetic resonance imaging in experienced healthy divers. Med Sci Monit. 2018;24:8279–89. https://doi.org/10.12659/MSM.911475.

118. Li Y, Jiang J, Shen T, Wu P, Zuo C. Radiomics features as predictors to distinguish fast and slow progression of mild cognitive impairment to Alzheimer's disease. Conf Proc IEEE Eng Med Biol Soc. 2018;2018:127–30. https://doi.org/10.1109/EMBC.2018.8512273.

119. Chitalia RD, Kontos D. Role of texture analysis in breast MRI as a cancer biomarker: a review. J Magn Reson Imaging. 2018;49(4):927–938. https://doi.org/10.1002/jmri.26556.

120. Sah BR, Owczarczyk K, Siddique M, Cook GJR, Goh V. Radiomics in esophageal and gastric cancer. Abdom Radiol (NY). 2018;44(6):2048-2058. https://doi.org/10.1007/s00261-018-1724-8.

121. Yang F, Ford JC, Dogan N, Padgett KR, Breto AL, et al. Magnetic resonance imaging (MRI)-based radiomics for prostate cancer radiotherapy. Transl Androl Urol. 2018;7(3):445–58. https://doi.org/10.21037/tau.2018.06.05.

122. Zhang X, Xu X, Tian Q, Li B, Wu Y, Yang Z, et al. Radiomics assessment of bladder cancer grade using texture features from diffusion-weighted imaging. J Magn Reson Imaging. 2017;46(5):1281–8. https://doi.org/10.1002/jmri.25669.

123. Rozenberg R, Thornhill RE, Flood TA, Hakim SW, Lim C, Schieda N. Whole-tumor quantitative apparent diffusion coefficient histogram and texture analysis to predict Gleason score upgrading in intermediate-risk 3 + 4 = 7 prostate cancer. AJR Am J Roentgenol. 2016;206(4):775–82. https://doi.org/10.2214/AJR.15.15462.

124. Nketiah G, Elschot M, Kim E, Teruel JR, Scheenen TW, Bathen TF, et al. T2-weighted MRI-derived textural features reflect

prostate cancer aggressiveness: preliminary results. Eur Radiol. 2017;27(7):3050–9. https://doi.org/10.1007/s00330-016-4663-1.

125. Sankar V, Kumar D, Clausi DA, Taylor GW, Wong A. SISC: end-to-end interpretable discovery radiomics-driven lung cancer prediction via stacked interpretable sequencing cells. https://arxiv.org/pdf/1901.04641

126. Pietikäinen M, Zhao G. Two decades of local binary patterns: a survey. In: Bingham E, Kaski S, Laaksonen J, Lampinen J (eds.) Advances in independent component analysis and learning machines. London: Academic; 2015. p. 175–210.

127. 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(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.

128. Ravì D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform. 2016;21(1):4–21.

129. Sharif Razavian A, Azizpour H, Sullivan J, Carlsson S. (2014). CNN features off-the-shelf: an astounding baseline for recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2014. p. 806–13).

130. Kawahara J, BenTaieb A, Hamarneh G. Deep features to classify skin lesions. In: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI); 2016. p. 1397–1400).

131. Merone M, Sansone C, Soda P. A computer-aided diagnosis system for HEp-2 fluorescence intensity classification. Artificial Intelligence in Medicine. 2019;97:71–78.

132. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn. 2011;3:1–122.

133. Karr AF, Lin X, Sanil AP, Reiter JP. Secure regression on distributed databases. J Comput Graph Stat. 2005;14:263–79. https://doi.org/10.1198/106186005X47714.

134. Karr AF, Lin X, Sanil AP, Reiter JP. Privacy-preserving analysis of vertically partitioned data using secure matrix products. J Off Stat. 2009;25:125.

135. Sanil AP, Karr AF, Lin X, Reiter JP. Privacy preserving regression modelling via distributed computation. In: Proc Tenth ACM SIGKDD Int Conf Knowl Discov Data Min ACM; 2004. p. 677–82.

136. Chen R, Sivakumar K, Kargupta H. Learning Bayesian network structure from distributed data. In: Barbara D, Kamath C (eds.) Proceedings of the 2003 SIAM International Conference on Data Mining (SDM 2003), San Francisco CA; 2003. p. 284–8.

137. Wright R, Yang Z. Privacy-preserving Bayesian network structure computation on distributed heterogeneous data. In: Proc Tenth ACM SIGKDD Int Conf Knowl Discov Data Min. New York, NY, USA: ACM; 2004. p. 713–8. https://doi.org/10.1145/1014052.1014145.

138. Yang Z, Wright RN. Improved privacy-preserving Bayesian network parameter learning on vertically partitioned data. 21st. Int Conf Data Eng Workshop. 2005;2005:1196. https://doi.org/10.1109/ICDE.2005.230.

139. Meng D, Sivakumar K, Kargupta H. Privacy-sensitive Bayesian network parameter learning. Data Min 2004 ICDM04 Fourth IEEE Int Conf On IEEE. 2004;2004:487–90.

140. Jochems A, Deist Timo M, van Soest J, Eble M, Bulens P, Coucke P, et al. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital — a real life proof of concept. Radiother Oncol. 2016;121:459–67.

141. Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, et al. Comparison of Bayesian network and support vector machine models for twoyear survival prediction in lung cancer patients treated with radiotherapy. Med Phys. 2010;37:1401–7.

142. Lualdi M, Fasano M. Statistical analysis of proteomics data: a review on feature selection. J Proteomics. 2018;198:18-26. https://doi.org/10.1016/j.jprot.2018.12.004.

143. Leijenaar RT, Carvalho S, Velazquez ER, van Elmpt WJ, 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(7):1391–7. https://doi.org/10.3109/0284186X.2013.812798.

144. O'Connor JP, Aboagye EO, Adams JE, Aerts HJ, Barrington SF, Beer AJ, et al. Imaging biomarker roadmap for cancer studies. Nat Rev Clin Oncol. 2017;14(3):169–86. https://doi.org/10.1038/nrclinonc.2016.162.

145. Aerts HJ. The potential of radiomic-based phenotyping in precision medicine: a review. JAMA Oncol. 2016;2(12):1636–42. https://doi.org/10.1001/jamaoncol.2016.2631.

146. Antunovic L, Gallivanone F, Sollini M, Sagona A, Invento A, Manfrinato G, et al. [18F]FDG PET/CT features for the molecular characterization of primary breast tumors. Eur J Nucl Med Mol Imaging. 2017;44(12):1945–54. https://doi.org/10.1007/s00259-017-3770-9.

147. El Naqa I, Li R, Murphy MJ. Machine learning in radiation oncology: theory and applications. SpringerLink; 2015.

148. Huang SS. Supervised feature selection: a tutorial. Artif Intell Res. 2015;4(2):22–37. https://doi.org/10.5430/air.v4n2p22.

149. Yoon HJ, Kim Y, Chung J, Kim BS. Predicting neo-adjuvant chemotherapy response and progression-free survival of locally advanced breast cancer using textural features of intratumoral heterogeneity on F-18 FDG PET/CT and diffusion-weighted MR imaging. Breast J. 2018;25(3):373–80.

150. Choi JW, Lee D, Hyun SH, Han M, Kim JH, Lee SJ. Intratumoural heterogeneity measured using FDG PET and MRI is associated with tumour–stroma ratio and clinical outcome in head and neck squamous cell carcinoma. Clin.Radiol. 2017;72:482–9.

151. Fave X, Mackin D, Yang J, Zhang J, Fried D, Balter P, et al. Can radiomics features be reproducibly measured from CBCT images for patients with non-small cell lung cancer? Med.Phys. 2015;42:6784–97.

152. Antunes J, Viswanath S, Rusu M, Valls L, Hoimes C, Avril N, et al. Radiomics analysis on FLT-PET/MRI for characterization of early treatment response in renal cell carcinoma: a proof-of-concept study. Transl.Oncol. 2016;9:155–62.

153. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Heidelberg: Springer; 2009

154. Fawcett T. ROC graphs: notes and practical considerations for researchers. Mach Learn. 2004;31(1):1–38.

155. Efron B, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hall/CRC;1994.

156. Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc. 1997;92(438):548–60.

157. Hawkins DM, Basak SC, Mills D. Assessing model fit by cross-validation. J Chem Inf Comput Sci. 2003;43(2):579–86.

158. Raschka S. Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808. 2018.

159. Vallieres M, Laberge S, Diamant A, El Naqa I. Enhancement of multimodality texture-based prediction models via optimization of PET and MR image acquisition protocols: a proof of concept. Phys.Med.Biol. 2017;62:8536–65.

160. Metz S, Ganter C, Lorenzen S, van Marwick S, Herrmann K, Lordick F, et al. Phenotyping of tumor biology in patients by multimodality multiparametric imaging: relationship of microcirculation, alphavbeta3 expression, and glucose metabolism. J Nucl Med. 2010;51:1691–8.

161. Soda P. A multi-objective optimisation approach for class imbalance learning. Pattern Recogn. 2011;44(8):1801–10.

162. Provost FJ, Fawcett T, Kohavi R. The case against accuracy estimation for comparing induction algorithms. ICML. 1998;98:445–53.

163. Efron B. Estimating the error rate of a prediction rule: improvement on cross-validation. J Am Stat Assoc. 1983;78(382):316–31.
164. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7(1):91.
165. Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. J Mach Learn Res. 2004;5:1089–105.
166. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Int Joint Conf Artif Intell. 1995;14(12):1137–43.
167. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21(15):3301–7.
168. Refaeilzadeh P, Tang L, Liu H. On comparison of feature selection algorithms. In: Proceedings of AAAI Workshop on Evaluation Methods for Machine Learning II; 2007. p. 34–9.
169. Avanzo M, Stancanello J, El Naqa I. Beyond imaging: the promise of radiomics. Phys.Med. 2017;38:122–39. https://doi.org/10.1016/j.ejmp.2017.05.071.
170. Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Trans Pattern Anal Mach Intell. 2000;22(1):4–38.
171. Lohmann P, Kocher M, Ceccon G, Bauer EK, Stoffels G, Viswanathan S, et al. Combined FET PET/MRI radiomics differentiates radiation injury from recurrent brain metastasis. Neuroimage Clin. 2018;20:537–42.
172. Vaidya M, Creach KM, Frye J, Dehdashti F, Bradley JD, El Naqa I. Combined PET/CT image characteristics for radiotherapy tumor response in lung cancer. Radiother Oncol. 2012;102(2):239–45.
173. El Naqa I, Suneja G, Lindsay PE, Hope AJ, Alaly JR, Vicic M, et al. Dose response explorer: an integrated open-source tool for exploring and modelling radiotherapy dose-volume outcome relationships. Phys Med Biol. 2006;51:5719–35.
174. Lucia F, Visvikis D, Desseroit MC, Miranda O, Malhaire JP, Robin P, et al. Prediction of outcome using pretreatment (18)F-FDG PET/CT and MRI radiomics in locally advanced cervical cancer treated with chemoradiotherapy. Eur J Nucl Med Mol Imaging. 2018;45:768–86.
175. Yin Q, Hung SC, Rathmell WK, Shen L, Wang L, Lin W, et al. Integrative radiomics expression predicts molecular subtypes of primary clear cell renal cell carcinoma. Clin Radiol. 2018;73:782–91.
176. Parmar C, Leijenaar RTH, Grossmann P, Velazquez ER, Bussink J, Rietveld D, et al. Radiomic feature clusters and prognostic signatures specific for lung and head & neck cancer. Sci Rep. 2015;5:11044.
177. Huang SY, Franc BL, Harnish RJ, Liu G, Mitra D, Copeland TP, et al. Exploration of PET and MRI radiomic features for decoding breast cancer phenotypes and prognosis. NPJ Breast Cancer. 2018;4:24.
178. Vallieres M, Freeman CR, Skamene SR, El Naqa I. A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys Med .Biol. 2015;60:5471–96.
179. Vriens D, Disselhorst JA, Oyen WJ, de Geus-Oei LF, Visser EP. Quantitative assessment of heterogeneity in tumor metabolism using FDG-PET. Int J Radiat Oncol Biol Phys. 2012;82:e725–31.
180. Crispin-Ortuzar M, Apte A, Grkovski M, Oh JH, Lee NY, Schoder H, et al. Predicting hypoxia status using a combination of contrast-enhanced computed tomography and [(18)F]-Fluorodeoxyglucose positron emission tomography radiomics features. Radiother Oncol. 2018;127:36–42.
181. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. https://arxiv.org/abs/1506.06579
182. Ganeshan B, Abaleke S, Young RCD, Chatwin CR, Miles KA. Texture analysis of non-small cell lung cancer on unenhanced computed tomography: initial evidence for a relationship with tumour glucose metabolism and stage. Cancer Imaging. 2010;10(1):137–43. https://doi.org/10.1102/1470-7330.2010.0021.
183. Strzelecki M, Szczypinski P, Materka A, Klepaczko A. A software tool for automatic classification and segmentation of 2D/3D medical image. Nucl Instr Meth Phys Res A. 2013;702:137–40. https://doi.org/10.1016/j.nima.2012.09.006.
184. Szczypinski P, Strzelecki M, Materka A, Klepaczko A. MaZda—A software package for image texture analysis. Comput Methods Prog Biomed. 2009;94(1):66–76. https://doi.org/10.1016/j.cmpb.2008.08.005.
185. Szczypinski P, Strzelecki M, Materka A. MaZda — a software for texture analysis. Proc. of ISITC 2007, November 23–23, 2007, Republic of Korea. 2007;2007:245–9.
186. 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(16):4786–9. https://doi.org/10.1158/0008-5472.CAN-18-0125.
187. Schaer R, Dicente Cid Y, Alkim E, John S, Rubin DL, Depeursinge A. Web-based tools for exploring the potential of quantitative imaging biomarkers in radiology: intensity and texture analysis on the ePAD platform. In: Biomedical texture analysis. London: Academic; 2017. p. 379–410. https://doi.org/10.1016/B978-0-12-812133-7.00013-2.
188. 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(21):e104–7. https://doi.org/10.1158/0008-5472.CAN-17-0339.
189. Dicente Cid Y, Castelli J, Schaer R, Scher N, Pomoni A, Prior O et al. QuantImage: an online tool for high-throughput 3D radiomics feature extraction in PET-CT. In: Biomedical texture analysis. London: Academic; 2017. p. 349–77. https://doi.org/10.1016/B978-0-12-812133-7.00012-0.
190. Zhang L, Fried DV, Fave XJ, Hunter LA, Yang J, Court LE. IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics. Med Phys. 2015;42(3):1341–53. https://doi.org/10.1118/1.4908210.
191. 3D Slicer — an open source software platform for medical image informatics, image processing, and three-dimensional visualization. https://www.slicer.org/.
192. 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(1):151–65.
193. Feedback Medical. https://fbkmed.com/
194. HealthMynehttps://www.healthmyne.com/
195. Quantib: AI in radiology. https://www.quantib.com/
196. Oncoradiomics SA. https://www.oncoradiomics.com/
197. Zebra Medical. https://www.zebra-med.com/
198. Median Technologies. http://www.mediantechnologies.com/
199. MaxQ AI. https://maxq.ai/
200. Arterys. https://www.arterys.com/
201. Imagia. https://www.imagia.com/
202. Regge D, Mazzetti S, Giannini V, Bracco C, Stasi M. Big data in oncologic imaging. Radiol Med. 2017;122(6):458–63. https://doi.org/10.1007/s11547-016-0687-5.
203. European Society of Radiology (ESR). ESR position paper on imaging biobanks. Insights imaging. 2015;6(4):403–10. https://doi.org/10.1007/s13244-015-0409-x.
204. Neri E, Regge D. Imaging biobanks in oncology: European perspective. Future Oncol. 2017;13(5):433–41. https://doi.org/10.2217/fon-2016-0239.
205. TheCancer Imaging Archive. http://www.cancerimagingarchive.net

206. UK Biobank. http://www.ukbiobank.ac.uk

207. German National Cohort (GNC) Consortium. The German National Cohort: aims, study design and organization. Eur J Epidemiol. 2014;29(5):371–82. https://doi.org/10.1007/s10654-014-9890-7.

208. Boyd S. Distributed optimization and statistical learning via the alternating direction method of multipliers. Found Trends Mach Learn. 2010;3(1):1–122.

209. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong JCC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol. 2017;14(12):749–62.

210. Clarke LP, Nordstrom RJ, Zhang H, Tandon P, Zhang Y, Redmond G. et al. The Quantitative Imaging Network: NCI's historical perspective and planned goals. Transl Oncol. 2014;7(1):1–4.

211. Cheng HT, Koc L, Harmsen J, Shaked T, Chandra T, Aradhye H, Anil R. Wide & deep learning for recommender systems. In: Workshop on Deep Learning for Recommender Systems; 2016. p. 7–10

212. Gonzalez ME, Dinelle K, Vafai N, Heffernan N, McKenzie J, Appel-Cresswell S, et al. Novel spatial analysis method for PET images using 3D moment invariants: applications to Parkinson's disease. Neuroimage. 2013;68:11–21. https://doi.org/10.1016/j.neuroimage.2012.11.055.

213. van Velden FH, Cheebsumon P, Yaqub M, Smit EF, Hoekstra OS, Lammertsma AA, et al. Evaluation of a cumulative SUV-volume histogram method for parameterizing heterogeneous intratumoural FDG uptake in non-small cell lung cancer PET studies. Eur J Nucl Med Mol Imaging. 2011;38:1636–47. https://doi.org/10.1007/s00259-011-1845-6.

214. Chawla NV, Bowyer KW, Hall LO, Philip Kegelmeyer W. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.

215. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. MedImage Anal. 2017;42:60–88.

216. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.

217. Teramoto A, Fujita H, Yamamuro O, Tamaki T. Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique. Med.Phys. 2016;43:2821–7.

218. Hosny, et al. Deep learning for lung cancer prognostication: a retrospective multi-cohort radiomics study. PLoS Med. 2018;15(11):e1002711.

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