Radiomics: Data Are Also Images

Mathieu Hatt[1], Catherine Cheze Le Rest[1,2], Florent Tixier[1], Bogdan Badic[1], Ulrike Schick[1], and Dimitris Visvikis[1]

[1]*LaTIM, INSERM, UMR 1101, Univ Brest, Brest, France; and* [2]*Nuclear Medicine Department, CHU Milétrie, Poitiers, France*

The aim of this review is to provide readers with an update on the state of the art, pitfalls, solutions for those pitfalls, future perspectives, and challenges in the quickly evolving field of radiomics in nuclear medicine imaging and associated oncology applications. The main pitfalls were identified in study design, data acquisition, segmentation, feature calculation, and modeling; however, in most cases, potential solutions are available and existing recommendations should be followed to improve the overall quality and reproducibility of published radiomics studies. The techniques from the field of deep learning have some potential to provide solutions, especially in terms of automation. Some important challenges remain to be addressed but, overall, striking advances have been made in the field in the last 5 y.

Key Words: radiomics; machine learning; deep learning

J Nucl Med 2019; 60:38S–44S
DOI: 10.2967/jnumed.118.220582

Nuclear medicine in oncology was revolutionized by the deployment of PET/CT (combined) scanners in the 2000s (*1*). Subsequent hardware and software innovations have led to improvements in the spatial resolution and the signal-to-noise ratio of reconstructed images, with point-spread-function and time-of-flight information being integrated into reconstruction algorithms (*2,3*). PET/CT images have always been quantitative, and their accuracy has been increased thanks to these improvements. However, PET/CT images have been exploited in a limited way in most clinical publications, clinical trials and, obviously, routine clinical practice. In most cases, nuclear medicine physicians detect and anatomically localize pathologic uptake visually. Subsequently, the identified lesions are characterized by a single semiquantitative parameter corresponding to the maximum-intensity voxel, known as the SUV_{max}. An aggregate of several voxels in a 1-cm^3 spheric region may be used (SUV_{peak}) to increase the robustness of the measurement with respect to statistical noise (*4*). Although the SUV_{max} has been successful in several clinical applications, including diagnosis and staging, it has also been shown to be insufficiently discriminative in several settings, such as baseline prognosis (*5*) or prediction of a response to therapy (*6*).

Nuclear medicine physicians need to go beyond such a simplistic metric, notwithstanding the fact that these data are also images. In that regard, the recent success of deep learning (DL) is a promising development, because DL is specifically aimed at learning patterns relevant for a given task (e.g., segmentation or endpoint prediction) from the data (i.e., images) themselves, instead of relying on "engineered" or "handcrafted" features (*7*).

In parallel to the improvements in hardware and reconstruction software, several developments in image processing, analysis, and machine learning have been applied to PET/CT and SPECT/CT. First, preprocessing algorithms such as denoising (*8,9*) and correction of partial-volume effects (*10*) have led to improvements in both qualitative and quantitative accuracy. Second, compared with experts, (semi)automated algorithms have been able to detect lesions of interest and delineate them with similar accuracy and higher reproducibility and robustness (*11*). Third, the extraction of quantitative metrics from PET and SPECT images to characterize tumors or organs of interest has been exponentially growing over the last 10 y, relying initially on engineered features (*12,13*) or, more recently, on "deep" features extracted using convolutional neural networks (CNNs) (*14*). Finally, the development of multiparametric models using machine learning for disease diagnosis or staging and predicting outcomes also has been exponentially increasing (*15,16*). These 4 methodologic foundations are key elements of the field of radiomics (*17,18*).

Radiomics considers images as quantitative data from which to extract information that may not be accessible to the naked eye, even the expertly trained one (*19*). Thus, "images are more than pictures, they are data" (*20*); however, images should not be forgotten—that is, data are also images. Although the content of an image can be reduced to a set of quantitative features, the entire image may still provide additional information; it is important to remember this fact with regard to the learning process of DL algorithms.

The goal of this commissioned article is to provide an update on the state of the art, pitfalls, solutions for those pitfalls, future perspectives, and challenges in the quickly evolving field of radiomics (i.e., images as data and vice versa) in nuclear medicine imaging and associated oncology applications.

PET AND SPECT RADIOMICS PUBLICATIONS

Although the radiomics approach was initially developed in the context of radiotherapy and radiology, the number of studies applying radiomics to PET or SPECT has been steadily increasing. On March 25, 2019, about 1,000 publications (excluding abstracts and meetings) using the term *radiomics* could be found in Web of Science databases—an exponential increase (Fig. 1). About 27% of them concerned PET or PET/CT, and only a few concerned SPECT/CT (e.g., (*21*)). However, almost one-quarter (22%) of them were editorials and reviews. Also, several papers published before or after the term was introduced could be considered "PET radiomics studies" (e.g., (*12,22–24*)).

Received Feb. 1, 2019; revision accepted Mar. 28, 2019.
For correspondence or reprints contact: Mathieu Hatt, LaTIM, INSERM, UMR 1101, IBRBS, Faculté de Médecine, 22 Rue Camille Desmoulins, 29238 Brest, France.
E-mail: hatt@univ-brest.fr

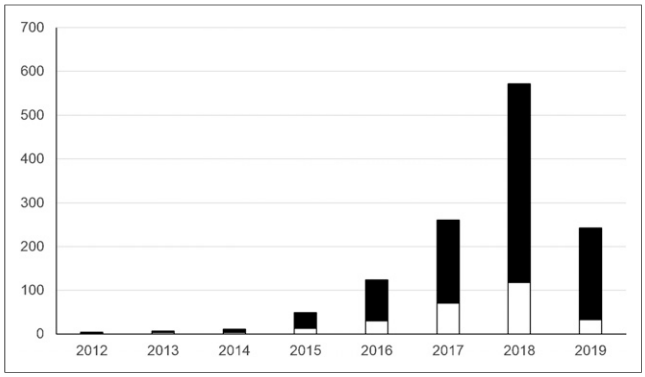

FIGURE 1. Evolution of number of publications found in Web of Science (all databases, black part) using the term *radiomics* and containing the term *PET* or *PET/CT* or *positron* (white part).

MAIN PITFALLS (AND SOLUTIONS FOR THOSE PITFALLS) IN NUCLEAR MEDICINE RADIOMICS STUDIES

The work flow of radiomics analysis is the same for any image modality and actually corresponds to the usual machine learning pipeline (Fig. 2): data (images) are input for an extractor (e.g., software calculating features), and then a modeling step is used to map the features to the classification goal (e.g., outcome for patients). This pipeline makes every step highly dependent on the methodologic choices made in the previous steps. Thus, there are several pitfalls in each of them.

Study Design

Before data collection is actually begun, it is important to define the question to be answered, to determine the kind (and quantity) of data needed to answer it, and to list the study needs and requirements. Several guidelines can help in the design of future studies (*25–28*), avoiding pitfalls typically associated with each of the next steps. For instance, we recommend relying on the radiomics quality score (*29*) and the TRIPOD (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis) guidelines (*30*). A specific example is to ensure having datasets of sufficient size and from different sources to satisfy the training, validation, and testing requirements (*31*).

Data Acquisition

Image Acquisition and Reconstruction. Data including images must be collected (retrospectively) or acquired (prospectively). When images are retrospectively collected, the associated raw data are not available; therefore, the reconstructed images must be exploited as they are. On the contrary, when images are acquired prospectively, raw data should be stored for research purposes or, at the very least, image reconstruction settings suitable for radiomics analyses should be chosen. Indeed, clinical reconstruction settings are usually optimized for visual analysis tasks that are mostly focused on detection rather than finer characterization—hence, larger voxel sizes (~4–5 mm, often nonisotropic) and postreconstruction smoothing of images with suboptimal gaussian filtering. For radiomics, smaller (facilitated delineation) and isotropic (unbiased texture computation) voxel sizes (Fig. 3) (*32*), without postfiltering, should be used—so that if the radiomics pipeline includes preprocessing steps (e.g., denoising, correction of partial-volume effects), these can be applied to unprocessed images.

Nonimage Data Collection. Collection of information from clinical records and other analyses (e.g., histopathology, transcriptomics, genetics) is a crucial step for which curation quality checks need to be provisioned in the study design. Indeed, this information is usually retrieved from medical records by investigators and manually entered into new research databases, a process prone to errors. Such errors introduced at this level of the work flow can be highly detrimental and complex to identify a posteriori, warranting the need for a well-designed data infrastructure (*25*).

Multicenter Data. The need for larger multicenter datasets was emphasized previously (*33,34*). Indeed, developing multiparametric models requires large, representative cohorts to train the models on relevant data and make them as clinically useful and generalizable as possible. Because sharing data in a single storage facility for a centralized analysis is complex for legal, ethical, administrative, and technical reasons, such sharing is not the reality of current radiomics

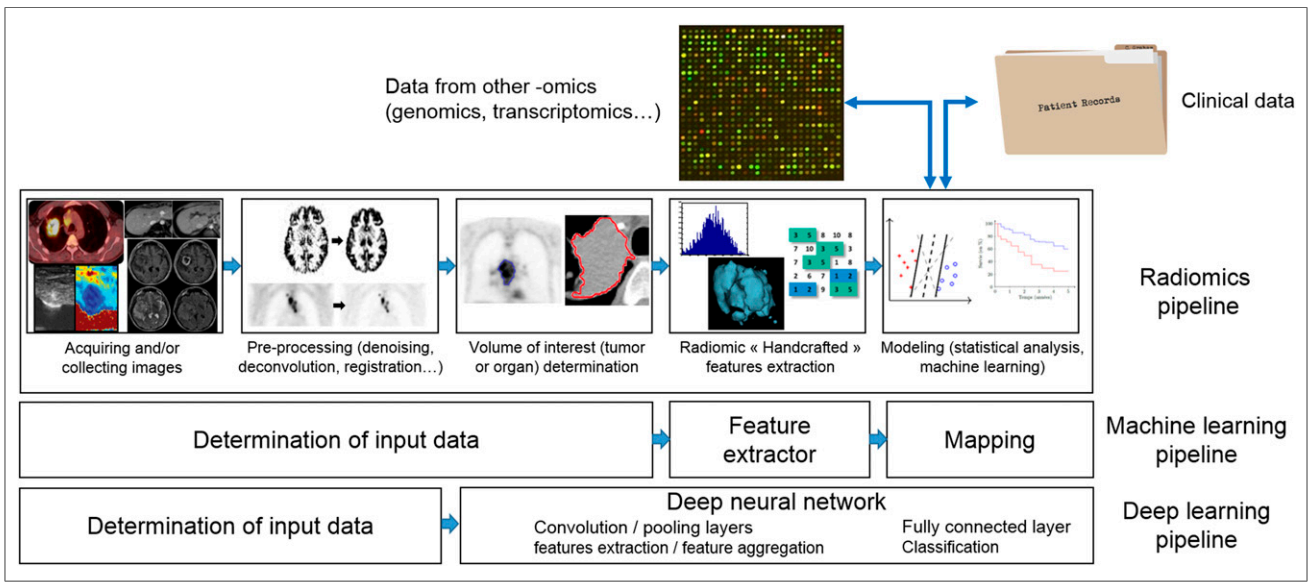

FIGURE 2. Radiomics pipeline in comparison with usual machine learning work flow.

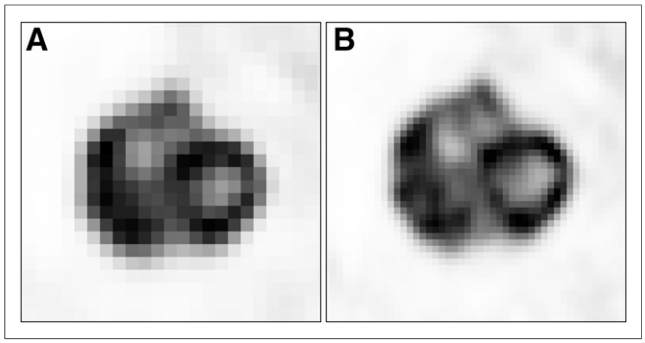

FIGURE 3. Axial slice of ^{18}F-FDG PET image of lung tumor reconstructed with 3-dimensional row-action maximization-likelihood algorithm using standard $4 \times 4 \times 4$ mm^3 (A) or finer $2 \times 2 \times 2$ mm^3 (B) voxels.

studies, especially in nuclear medicine. As a result, most published radiomics models have not been properly validated (*35*).

Distributed learning provides a solution to train a model at each institution and update the parameters of the model in a centralized computing station without the data ever leaving clinical centers, as only the parameters of the model are exchanged (*36*). However, whether images are processed locally or in a centralized fashion, differences in image properties and the resulting variability of features need to be taken into account to build robust models. Indeed, several studies showed that most radiomic features are sensitive (to a variable degree) to differences in scanner models, acquisition parameters, and reconstruction settings (*23,37*).

Several options for addressing these issues are available. First, standardizing PET/CT acquisition and reconstruction protocol settings is an important aspect of multicenter data collection, with guidelines already available for PET/CT imaging (*38,39*). However, these are still mainly focused on the SUV and do not include standardization recommendations regarding radiomics, for which harmonization may be more difficult to achieve. Although these efforts should be expanded for radiomics (*40*) and can help reduce differences in radiomic feature distributions across different sites, they may not be sufficient. In addition, this approach is feasible only for prospectively collected data, but most radiomics studies are still performed by retrospectively analyzing available data.

Second, preprocessing images can help reduce differences, for example, by interpolating them to a common voxel size and filtering so that they have similar resolution and noise characteristics. This approach may be insufficient to suppress all differences in the resulting radiomic feature distributions. Also, performing this approach is not trivial, as there are dozens of algorithms for interpolating and filtering, and identifying an optimal combination is challenging. This approach can also introduce artifacts or reduce the quality of the quantitative information contained in images.

Removing features identified as too sensitive to the variability of acquisition and reconstruction settings is another solution that can help build more robust models when used with an external dataset. This solution has been extensively studied for several modalities, including PET (*23,41–44*), but similar studies for SPECT are currently lacking. Most of these studies have shown that the robustness, repeatability, or reliability of features (with respect to acquisition, reconstruction, filtering, segmentation, or analysis and computation choices) is highly variable among features in general as well as among features of a given category or specific matrix. The main drawback of this approach is the potential loss of information, as numerous features will be discarded even though they may contain clinically relevant information. Another limitation is that identifying the optimal subset of features that is sufficiently robust and provides enough discriminative power is challenging and likely must be done for each combination of image modality, cancer type, and task. A recently evaluated method consists of dealing with the variability of radiomic features from each dataset or cohort a posteriori in the modeling step itself, by harmonizing feature distributions so that they can be pooled together.

Several methods have been developed to address the same issues in genomics, for which the "batch effect" has a significant impact. The ComBat method (*45*) has been shown to work well for small samples and to outperform similar techniques (*46*). ComBat was shown to allow PET radiomics predictive models to achieve higher performance in the external validation step (*47,48*). This approach has several advantages: it is easy and fast, and it allows all of the information to be exploited because all of the features are retained. One limitation is that a sample population of at least 30 patients from each center dataset must be available, and it cannot be applied on an individual-patient basis. Other techniques, such as rescaling and normalization, were recently evaluated for improving multicenter modeling, with interesting results (*49*).

Segmentation

The delineation of the object of interest (e.g., a tumor) is the most time-consuming bottleneck step, as full automation is difficult to achieve. In most studies, an expert first isolates the object of interest in a manually or semiautomatically defined volume of interest, and then a (semi)automated algorithm is used for actually delineating the object. This step is more complex for diffuse disease or several lesions. PET tumor delineation has been investigated in numerous studies (*11*). Despite perfect repeatability and very high interobserver reproducibility, threshold-based techniques have been shown to perform poorly in terms of robustness and absolute accuracy, especially for heterogeneous uptake distributions (*11*). Manual delineation has well-known limitations regarding inter- and intraobserver variability and should be performed by at least 2 (preferably more) experts, with consensus. A recent MICCAI (Medical Image Computing and Computer-Assisted Intervention) challenge highlighted the poor performance of fixed thresholding as well as the ability of more advanced techniques to achieve higher accuracy (*50*).

An alternative for reducing variability in the performance of individual algorithms and obtaining a more consistent result across a given dataset is to consider the statistical consensus of several methods (*51*). Another potential solution is to train an algorithm to select the best method for a given case depending on image properties and other a priori information (*52*). DL has been especially successful in medical image segmentation tasks (*7*), as the learning process occurs on the voxel level and not on the entire-image level (as for classification tasks), thereby reducing the requirements regarding the amount of learning data needed for efficient training. Recently, convolutional neural network approaches were applied to PET (*50*) and PET/CT segmentation (*53–55*). DL algorithms for PET tumor detection and segmentation (*56*) may provide fully automated solutions for these steps of the radiomics pipeline. Similar efforts have been made to characterize disease in PET/CT images without the use of DL methods (*57*).

Feature Calculation

Standardization and Nomenclature. The main pitfall related to this step is the lack of standardization of both nomenclature and

implementation. The calculation of features involves several steps and several different choices (mostly for textural features); their implementation is therefore prone to errors.

With regard to these issues, the efforts of the Imaging Biomarkers Standardization Initiative (IBSI) (*26,58*) should be emphasized. This initiative is performed by more than 20 research groups from 8 countries and aims to establish standardized definitions of usual radiomic features (currently 172) and their calculation; a common nomenclature for the full radiomics pipeline and each step of pre- and postprocessing leading to feature extraction; recommendations regarding interpolation, discretization, and texture matrix design; a benchmark of standardized values based on both a synthetic digital phantom and real clinical images for each radiomic feature calculated in different possible configurations; and recommendations regarding reporting. The 123-page reference document (version 9, updated May 2019) is available online and published as a preprint (*26*). Although the IBSI is not specifically dedicated to PET imaging, most of its recommendations and results are directly applicable to PET radiomics studies. For instance, we highly recommend checking the IBSI compliance of homemade or commercial/open-source libraries and software before using them in a study, as doing so will greatly increase its reproducibility (*27*).

Confounding Issues for Volume and Other Metrics. Radiomic features include standard PET metrics (e.g., functional volume, mean, or SUV_{max}). Regarding additional, more complex quantitative measurements, such as geometric descriptors (e.g., sphericity or surface irregularity) or second- and higher-order textural features (e.g., $entropy_{GLCM}$ or $GLNU_{GLRLM}$), it is important to check their redundancy and complementary values with both clinical factors (e.g., stage or sex) and other available variables as well as standard PET metrics (e.g., volume or SUV_{max}). It is pointless to calculate complex image features that are simply surrogates of these (*41*). This issue is especially important for metabolic volume, as all radiomic features are calculated from a previously determined volume of interest through segmentation of the tumor. It has been shown that the design choices made in the calculation of features, such as the method and parameters used in the discretization of original intensities or the merging strategies of texture matrices, can have a tremendous impact on feature distributions and correlative relationships with volume or SUV_{max} (*33,43,44,59,60*). For example, regarding PET radiomics, it was shown by Hatt et al. (*41*) that the textural features previously identified by Tixier et al. (*24*) to predict a response to chemoradiotherapy in esophageal cancer were actually highly correlated with the corresponding volume and therefore provided little to no additional information, with a predictive ability similar to that of the volume alone. It was later shown that through different calculation settings (discretization and texture matrix design), the same textural features can provide complementary or additional value relative to volume, including for small tumors. Thus, their combination could lead to better stratification of patients (*44*), contrary to previous claims that no such complementary value could be obtained for volumes of less than 45 cm^3 (*61*).

Another metric, the so-called "heterogeneity factor"—defined as the derivative (dV/dT) of the volume–threshold function—was reported to be highly correlated with functional volume (*62*) and was therefore a surrogate of volume rather than an actual heterogeneity measurement (*63*). Similarly, the CT-derived radiomics signature for lung and head and neck cancers (*64*) was demonstrated to actually reflect mostly tumor volume rather than actual tumor heterogeneity and shape, as the shape (compactness) and textural ($GLNU_{GLRLM}$) parameters selected for this signature were later shown to be highly correlated with the corresponding volume (*65,66*). However, it was also shown that by adopting modified feature definitions, as proposed in the IBSI (i.e., dimensionless, compact, normalized, and merged textural matrices for $GLNU_{GLRLM}$ calculation), it was possible to obtain a higher prognostic power of the same signature compared with volume (*65*).

New Features. A single feature could have a large number of different values according to the choice of various parameters, including—but not limited to—the intensity discretization method and parameter(s) or merging strategy (directions and averaging). Although the feature definition is always the same, the obtained value can vary greatly from one matrix design to the next and therefore can create an additional variable for the analysis. This variability can actually be a way to optimize texture analysis, as each feature may end up being more informative with specific and different calculation choices (*32,60*). The robustness of features with respect to their calculation compared with their clinical discriminative power should not be overemphasized; further investigation to determine which features are indeed robust enough with respect to their level of discriminative power for a given endpoint is warranted (*67*).

New "engineered" or "handcrafted" features with potentially higher discriminative power or better properties are continuously being developed. CoLIAGe (Cooccurrence of Local Anisotropic Gradient Orientations) (*68*), a metabolic gradient (*69*), or 3-dimensional Riesz-covariance textures (*70*) are examples of such new features with a potentially higher differentiation power compared with standard textural features. A novel metric for quantifying PET heterogeneity was also proposed as a more intuitive and simple alternative to textural features; this method involves summing voxelwise distributions of differential SUVs, weighted by the distance of SUV differences among neighboring voxels from the center of the tumor (*71*). This metric was designed to yield increased values for tumors with peripheral subregions having high SUVs. A new grey-level cooccurrence matrix methodology was recently developed to reduce the redundancy of resulting features, demonstrating a more accurate classification of tumor types in CT images (*72*). Even if some of these metrics were not specifically developed for PET imaging, they could be directly applied to PET.

Finally, DL has also been the source of new features, commonly denoted as "deep features." These can be extracted from medical images using pretrained networks. These networks may have been trained on very large medical image datasets as well as on natural images (*73*). Because these networks have learned from natural images to extract rough to finer features at different scales through different layers, they can extract similar patterns and features from medical images (including PET) and can be used "off the shelf" or after an additional fine-tuning step (also called transfer learning). Most current results obtained with deep features as well as their combination with typical radiomic features have been obtained in CT and MRI applications (*74–78*), but the same concept can be applied to PET.

Modeling

Statistical analysis for mapping the extracted features to a given endpoint (either classification or regression) is one of the most challenging steps in the entire radiomics process. The goal of this step is usually to identify the optimal combination of the fewest available variables (clinical data, radiomics, and other analyses) allowing the maximization of 1 or several criteria (usually the

receiver operating characteristic area under the curve, concordance statistic, specificity, sensitivity, or accuracy). Indeed, statistical analysis was the weakest part of most texture and radiomics studies before 2015 because it tested too many hypotheses (i.e., number of features) for small patient cohorts without correction for type I errors (i.e., false discovery) and without the use of a validation dataset, thereby reporting mere (overfitted) correlations and not actual predictive power. Most radiomics or texture studies with PET have been performed with cohorts of fewer than 150 patients (*48*) and—because the number of features (and variables) is constantly growing, especially in the case of texture optimization (i.e., calculation of each feature with different parameters)—statistical analysis is fraught with the curse of dimensionality, a high rate of false-positive results, collinearity issues, and risk of overfitting. Choosing a machine learning method is also quite challenging. The most recent comparison studies highlighted the differences between popular methods as well as the fact that none of them performed best across the entire spectrum of datasets and tasks (*15,16,79*).

Following simple guidelines for robust and reliable statistical analysis and machine learning (*31*) is crucially important for obtaining reproducible and reliable results. The most important guidelines are splitting the available data into a training set (i.e., learning a model, e.g., a linear combination of 2 variables) and a validation set (i.e., tuning the parameters of the model, e.g., the weights of each variable in the linear combination) and performing the final evaluation with a testing set (i.e., performing the trained model with fixed parameters using a dataset never used in the training and validation steps). In the context of radiomics, different strategies can be used. Splitting a single available dataset into 3 parts is a potential solution. For example, a 100-patient cohort can be split into a training set of 50 patients, a validation set of 30 patients, and a final, testing set of 20 patients. Obviously, the larger the cohort, the better, as evaluating the final model with only 20 patients can provide limited evidence of its usefulness. Alternatively, if different datasets are available (e.g., in a multicenter setting), then it may be appropriate to train and validate with 1 cohort and test the resulting model with other cohorts (*47,80*). However, this approach requires harmonization of the features because of differences in their distributions across centers.

Different techniques can be used for splitting; we recommend either using stratified sampling (*81*) to ensure similar distributions in the splits or performing several different splits randomly and reporting the mean and SD for the results. Indeed, random splits can lead to very different distributions in the training, validation, and testing sets (e.g., all "easy" cases end up in the training set or, worse, the training set contains all of the cases to detect but the testing set contains none).

Another important pitfall concerns the imbalance of the data and classification (or regression) task, combined with the metrics used for performance evaluation. In most radiomics studies, the clinical endpoint is not balanced; that is, 1 class (e.g., patients with recurrence) dominates the other (e.g., patients without recurrence). In such a context, a machine learning algorithm classifying each instance as the dominating class would end up being right most of the time. Therefore, it is important to implement strategies to help an algorithm learn the minority cases as well as the majority cases, despite having fewer training examples. Several strategies are available; these include synthesizing additional minority instances (e.g., with the Synthetic Minority Oversampling Technique, [SMOTE] (*82*)), oversampling the minority class, undersampling the majority class, or tweaking the function cost to raise the cost of minority instance misclassification. Furthermore, it is important to rely on appropriate performance metrics, especially in the case of imbalanced data. For example, the often-considered metrics accuracy or F_1 score can indeed provide a biased estimation in the case of imbalanced data; the use of balanced accuracy (the mean of sensitivity and specificity), receiver operating characteristic area under the curve, and Matthews correlation coefficients is recommended instead to provide a reliable estimate of the performance of the model (*31*). For survival analysis and regression tasks, the use of hazard ratios and the concordance statistic is appropriate for evaluating time-to-event endpoints (*15*).

Finally, as there seem to be no currently available classifiers or feature selection methods that perform best across the entire spectrum of tasks and types of data, it may be interesting to consider ensemble techniques and the fusion of several different classifiers as a way to obtain more robust models (*83*).

REMAINING CHALLENGES TO ADDRESS AND HOW TO MOVE TO CLINICAL TRANSLATION

To enable clinical translation, despite numerous recent efforts, the radiomics community still has to address the following main challenges to help reduce the current limitations for both robust and reproducible research as well as actual clinical transfer: finalize and expand standardization efforts; develop tools and methods for collecting, storing, and sharing sufficiently large databases containing images associated with contextual clinical data and other analyses for a large panel of pathologies; reach a level of full automation for the entire pipeline (especially for the detection and segmentation steps); and identify and standardize optimal methods for model building and validation.

Regarding the collection of larger datasets, the main limitation preventing multicenter studies from reaching their full potential (i.e., the sensitivity of most radiomic features to variations in scanner models, acquisition protocols, and reconstruction settings) can be considered resolved on the basis of the use of a posteriori harmonization methods (*45,47,49*). For most of the remaining challenges, DL techniques can provide potential solutions, either for each of the steps in the radiomics work flow or by entirely replacing the usual work flow with an end-to-end DL-based approach (*14,84*). In the latter approach, all steps performed separately and sequentially (segmentation, feature extraction, modeling) are now performed by 1 (or several) neural network(s). This approach mostly replaces previous challenges with others specific to the use of DL techniques, such as the need for datasets much larger than those usually available in radiomics studies for efficient training. Therefore, techniques such as transfer learning and data augmentation become crucially important. Another requirement is to provide interpretable models by opening the "black box" that such networks, with the millions of parameters they contain, can appear to be. This requirement could be met by network visualization techniques (*85*), providing some visual feedback to end users and explaining why and how the network reached its final prediction—for instance, by providing heat maps on the original input images to highlight the most relevant areas in the images or even within the tumor.

CONCLUSION

The field of radiomics has been exponentially growing, including in PET/CT imaging. It is a very active and promising field of research, but it is full of methodologic pitfalls. Until recently, the approach has

been mostly to consider images as data by reducing full 3-dimensional images to a vector containing relevant quantitative handcrafted radiomic features. With the advent of DL techniques to solve challenges and lift the limitations of the current radiomics work flow, the radiomics community is returning to images as a whole; in this approach, patterns are captured by multilayer neuronal networks that learn the relevant features instead of selection and combination of handcrafted features.

DISCLOSURE

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Beyer T, Townsend DW, Brun T, et al. A combined PET/CT scanner for clinical oncology. *J Nucl Med.* 2000;41:1369–1379.
2. Berg E, Cherry SR. Innovations in instrumentation for positron emission tomography. *Semin Nucl Med.* 2018;48:311–331.
3. Jones T, Townsend D. History and future technical innovation in positron emission tomography. *J Med Imaging (Bellingham).* 2017;4:011013.
4. 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 1):122S–150S.
5. Bonomo P, Merlotti A, Olmetto E, et al. What is the prognostic impact of FDG PET in locally advanced head and neck squamous cell carcinoma treated with concomitant chemo-radiotherapy? A systematic review and meta-analysis. *Eur J Nucl Med Mol Imaging.* 2018;45:2122–2138.
6. Kwee RM. Prediction of tumor response to neoadjuvant therapy in patients with esophageal cancer with use of ^{18}F FDG PET: a systematic review. *Radiology.* 2010;254:707–717.
7. Litjens G, Kooi T, Bejnordi BE, et al. A survey on deep learning in medical image analysis. *Med Image Anal.* 2017;42:60–88.
8. Le Pogam A, Hanzouli H, Hatt M, Cheze Le Rest C, Visvikis D. Denoising of PET images by combining wavelets and curvelets for improved preservation of resolution and quantitation. *Med Image Anal.* 2013;17:877–891.
9. Gong K, Guan J, Liu C-C, Qi J. PET image denoising using a deep neural network through fine tuning. *IEEE Trans Radiat Plasma Med Sci.* 2019;3:153–161.
10. Le Pogam A, Hatt M, Descourt P, et al. Evaluation of a 3D local multiresolution algorithm for the correction of partial volume effects in positron emission tomography. *Med Phys.* 2011;38:4920–4923.
11. Hatt M, Lee JA, Schmidtlein CR, et al. Classification and evaluation strategies of auto-segmentation approaches for PET: report of AAPM Task Group No. 211. *Med Phys.* 2017;44:e1–e42.
12. El Naqa I, Grigsby P, Apte A, et al. Exploring feature-based approaches in PET images for predicting cancer treatment outcomes. *Pattern Recognit.* 2009;42:1162–1171.
13. Rahmim A, Salimpour Y, Jain S, et al. Application of texture analysis to DAT SPECT imaging: relationship to clinical assessments. *Neuroimage Clin.* 2016;12:e1–e9.
14. Ypsilantis P-P, Siddique M, Sohn H-M, et al. Predicting response to neoadjuvant chemotherapy with PET imaging using convolutional neural networks. *PLoS One.* 2015;10:e0137036.
15. 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.
16. Parmar C, Grossmann P, Bussink J, Lambin P, Aerts HJWL. Machine learning methods for quantitative radiomic biomarkers. *Sci Rep.* 2015;5:13087.
17. Gillies RJ, Anderson AR, Gatenby RA, Morse DL. The biology underlying molecular imaging in oncology: from genome to anatome and back again. *Clin Radiol.* 2010;65:517–521.
18. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* 2012;48:441–446.
19. Hatt M, Tixier F, Visvikis D, Cheze Le Rest C. Radiomics in PET/CT: more than meets the eye? *J Nucl Med.* 2017;58:365–366.
20. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology.* 2016;278:563–577.
21. Rahmim A, Huang P, Shenkov N, et al. Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images. *Neuroimage Clin.* 2017;16:539–544.
22. O'Sullivan F, Roy S, Eary J. A statistical measure of tissue heterogeneity with application to 3D PET sarcoma data. *Biostatistics.* 2003;4:433–448.
23. Galavis PE, Hollensen C, Jallow N, Paliwal B, Jeraj R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol.* 2010;49:1012–1016.
24. Tixier F, Le Rest CC, Hatt M, et al. Intratumor heterogeneity characterized by textural features on baseline ^{18}F-FDG PET images predicts response to concomitant radiochemotherapy in esophageal cancer. *J Nucl Med.* 2011;52:369–378.
25. Morin O, Vallières M, Jochems A, et al. A deep look into the future of quantitative imaging in oncology: a statement of working principles and proposal for change. *Int J Radiat Oncol Biol Phys.* 2018;102:1074–1082.
26. Zwanenburg A, Leger S, Vallières M, Löck S; for the Image Biomarker Standardisation Initiative. Image Biomarker Standardisation Initiative. https://arxiv.org/abs/1612.07003. Updated May 16, 2019. Accessed May 9, 2019.
27. Vallières M, Zwanenburg A, Badic B, Cheze Le Rest C, Visvikis D, Hatt M. Responsible radiomics research for faster clinical translation. *J Nucl Med.* 2018;59:189–193.
28. Aerts HJWL. Data science in radiology: a path forward. *Clin Cancer Res.* 2018;24:532–534.
29. Sanduleanu S, Woodruff HC, de Jong EEC, et al. Tracking tumor biology with radiomics: a systematic review utilizing a radiomics quality score. *Radiother Oncol.* 2018;127:349–360.
30. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD Statement. *BMJ.* 2015;350:g7594.
31. Chicco D. Ten quick tips for machine learning in computational biology. *BioData Min.* 2017;10:35.
32. 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;60:5471–5496.
33. Hatt M, Tixier F, Pierce L, Kinahan PE, Le Rest CC, Visvikis D. Characterization of PET/CT images using texture analysis: the past, the present ... any future? *Eur J Nucl Med Mol Imaging.* 2017;44:151–165.
34. O'Connor JPB, Aboagye EO, Adams JE, et al. Imaging biomarker roadmap for cancer studies. *Nat Rev Clin Oncol.* 2017;14:169–186.
35. Zwanenburg A, Löck S. Why validation of prognostic models matters. *Radiother Oncol.* 2018;127:370–373.
36. Jochems A, Deist TM, van Soest J, 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;121:459–467.
37. Yan J, Chu-Shern JL, Loi HY, et al. Impact of image reconstruction settings on texture features in ^{18}F-FDG PET. *J Nucl Med.* 2015;56:1667–1673.
38. Boellaard R, Delgado-Bolton R, Oyen WJG, et al. FDG PET/CT: EANM procedure guidelines for tumour imaging—version 2.0. *Eur J Nucl Med Mol Imaging.* 2015;42:328–354.
39. Kaalep A, Sera T, Rijnsdorp S, et al. Feasibility of state of the art PET/CT systems performance harmonisation. *Eur J Nucl Med Mol Imaging.* 2018;45:1344–1361.
40. Aide N, Lasnon C, Veit-Haibach P, Sera T, Sattler B, Boellaard R. EANM/EARL harmonization strategies in PET quantification: from daily practice to multicentre oncological studies. *Eur J Nucl Med Mol Imaging.* 2017;44(suppl 1):17–31.
41. Hatt M, Tixier F, Cheze Le Rest C, Pradier O, Visvikis D. Robustness of intratumour ^{18}F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma. *Eur J Nucl Med Mol Imaging.* 2013;40:1662–1671.
42. 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 ^{18}F-FDG PET. *J Nucl Med.* 2012;53:693–700.
43. Desseroit M-C, Tixier F, Weber WA, et al. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non–small cell lung cancer tumors: a repeatability analysis in a prospective multicenter cohort. *J Nucl Med.* 2017;58:406–411.
44. Hatt M, Majdoub M, Vallières M, et al. ^{18}F-FDG PET uptake characterization through texture analysis: investigating the complementary nature of heterogeneity and functional tumor volume in a multi–cancer site patient cohort. *J Nucl Med.* 2015;56:38–44.
45. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8:118–127.
46. Chen C, Grennan K, Badner J, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One.* 2011;6:e17238.
47. Lucia F, Visvikis D, Vallières M, et al. External validation of a combined PET and MRI radiomics model for prediction of recurrence in cervical cancer

patients treated with chemoradiotherapy. *Eur J Nucl Med Mol Imaging.* 2019;46: 864–877.
48. Orlhac F, Boughdad S, Philippe C, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med.* 2018;59:1321–1328.
49. Chatterjee A, Vallières M, Dohan A, et al. Creating robust predictive radiomic models for data from independent institutions using normalization. *IEEE Trans Radiat Plasma Med Sci.* 2019;3:210–215.
50. Hatt M, Laurent B, Ouahabi A, et al. The first MICCAI challenge on PET tumor segmentation. *Med Image Anal.* 2018;44:177–195.
51. McGurk RJ, Bowsher J, Lee JA, Das SK. Combining multiple FDG-PET radiotherapy target segmentation methods to reduce the effect of variable performance of individual segmentation methods. *Med Phys.* 2013;40:042501.
52. Berthon B, Marshall C, Evans M, Spezi E. ATLAAS: an automatic decision tree-based learning algorithm for advanced image segmentation in positron emission tomography. *Phys Med Biol.* 2016;61:4855–4869.
53. Zhao X, Li L, Lu W, Tan S. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. *Phys Med Biol.* 2018;64:015011.
54. Zhong Z, Kim Y, Plichta K, et al. Simultaneous cosegmentation of tumors in PET-CT images using deep fully convolutional networks. *Med Phys.* 2019;46: 619–633.
55. Guo Z, Li X, Huang H, Guo N, Li Q. Deep learning-based image segmentation on multi-modal medical imaging. *IEEE Trans Radiat Plasma Med Sci.* 2019;1–1.
56. Blanc-Durand P, Van Der Gucht A, Schaefer N, Itti E, Prior JO. Automatic lesion detection and segmentation of ^{18}F-FET PET in gliomas: a full 3D U-Net convolutional neural network study. *PLoS One.* 2018;13:e0195798.
57. Tong Y, Udupa JK, Odhner D, Wu C, Schuster SJ, Torigian DA. Disease quantification on PET/CT images without explicit object delineation. *Med Image Anal.* 2019;51:169–183.
58. Hatt M, Vallieres M, Visvikis D, Zwanenburg A. IBSI: an international community radiomics standardization initiative [abstract]. *J Nucl Med.* 2018;59(suppl 1): 287.
59. Leijenaar RTH, Nalbantov G, Carvalho S, 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.
60. Vallières M, Kay-Rivest E, Perrin LJ, et al. Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. *Sci Rep.* 2017;7:10117.
61. Brooks FJ, Grigsby PW. The effect of small tumor volumes on studies of intratumoral heterogeneity of tracer uptake. *J Nucl Med.* 2014;55:37–42.
62. Son SH, Kim D-H, Hong CM, et al. Prognostic implication of intratumoral metabolic heterogeneity in invasive ductal carcinoma of the breast. *BMC Cancer.* 2014;14:585.
63. Brooks FJ, Grigsby PW. Current measures of metabolic heterogeneity within cervical cancer do not predict disease outcome. *Radiat Oncol.* 2011;6:69.
64. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun.* 2014; 5:4006.
65. Vallieres M, Visvikis D, Hatt M. Dependency of a validated radiomics signature on tumor volume and potential corrections [abstract]. *J Nucl Med.* 2018;59(suppl 1):640.
66. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. *Radiother Oncol.* 2019;130:2–9.
67. Lv W, Yuan Q, Wang Q, et al. Robustness versus disease differentiation when varying parameter settings in radiomics features: application to nasopharyngeal PET/CT. *Eur Radiol.* 2018;28:3245–3254.
68. Prasanna P, Tiwari P, Madabhushi A. Co-occurrence of Local Anisotropic Gradient Orientations (CoLIAGe): a new radiomics descriptor. *Sci Rep.* 2016;6: 37241.
69. Wolsztynski E, O'Sullivan F, Keyes E, O'Sullivan J, Eary JF. Positron emission tomography-based assessment of metabolic gradient and other prognostic features in sarcoma. *J Med Imaging (Bellingham).* 2018;5:024502.
70. Cirujeda P, Dicente Cid Y, Muller H, et al. A 3-D Riesz-covariance texture model for prediction of nodule recurrence in lung CT. *IEEE Trans Med Imaging.* 2016;35:2620–2630.
71. Wang P, Xu W, Sun J, et al. A new assessment model for tumor heterogeneity analysis with ^{18}F-FDG PET images. *EXCLI J.* 2016;15:75–84.
72. Li X, Guindani M, Ng CS, Hobbs BP. Spatial Bayesian modeling of GLCM with application to malignant lesion characterization. *J Appl Stat.* 2019;46:230–246.
73. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis.* 2015;115:211–252.
74. Paul R, Hawkins SH, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. *J Med Imaging (Bellingham).* 2018;5:011021.
75. Ning Z, Luo J, Li Y, et al. Pattern classification for gastrointestinal stromal tumors by integration of radiomics and deep convolutional features. *IEEE J Biomed Health Inform.* 2019;23:1181–1191.
76. Antropova N, Huynh BQ, Giger ML. A deep feature fusion methodology for breast cancer diagnosis demonstrated on three imaging modality datasets. *Med Phys.* 2017;44:5162–5171.
77. Lao J, Chen Y, Li Z-C, et al. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep.* 2017;7:10353.
78. Li Z, Wang Y, Yu J, Guo Y, Cao W. Deep learning based radiomics (DLR) and its usage in noninvasive IDH1 prediction for low grade glioma. *Sci Rep.* 2017;7: 5467.
79. Deist TM, Dankers FJWM, Valdes G, et al. Machine learning algorithms for outcome prediction in (chemo)radiotherapy: an empirical comparison of classifiers. *Med Phys.* 2018;45:3449–3459.
80. Lucia F, Visvikis D, Desseroit M-C, 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;45:768–786.
81. Särndal C-E, Swensson B, Wretman J. *Model Assisted Survey Sampling.* New York, NY: Springer-Verlag; 1992.
82. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res.* 2002;16:321–357.
83. Paul R, Hall L, Goldgof D, Schabath M, Gillies R. Predicting nodule malignancy using a CNN ensemble approach. *Proc Int Jt Conf Neural Netw.* October 15, 2018 [Epub ahead of print].
84. Amyar A, Ruan S, Gardin I, Chatelain C, Decazes P, Modzelewski R. 3-D RPET-NET: development of a 3-D PET imaging convolutional neural network for radiomics analysis and outcome prediction. *IEEE Trans Radiat Plasma Med Sci.* 2019;3:225–231.
85. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H. Understanding neural networks through deep visualization. In: *Deep Learning Workshop, 31st International Conference on Machine Learning*; July 10–11, 2015; Lille, France.