European Radiology (2023) 33:7556–7560
https://doi.org/10.1007/s00330-023-09768-w

European Radiology
EUROPEAN SOCIETY OF RADIOLOGY

EDITORIAL

Assessment of RadiomIcS rEsearch (ARISE): a brief guide for authors, reviewers, and readers from the Scientific Editorial Board of *European Radiology*

Burak Kocak[1] · Leonid L. Chepelev[2] · Linda C. Chu[3] · Renato Cuocolo[4] · Brendan S. Kelly[5,6] · Philipp Seeböck[7] · Yee Liang Thian[8] · Robbert W. van Hamersvelt[9] · Alan Wang[10] · Stuart Williams[11] · Jan Witowski[12] · Zhongyi Zhang[13] · Daniel Pinto dos Santos[14,15]

Received: 7 March 2023 / Revised: 24 March 2023 / Accepted: 14 April 2023 / Published online: 26 June 2023

Introduction

A simple *PubMed* search using the term "radiomics" on February 28, 2023, reveals 7857 publications at the time of the writing, with no time filter applied. Despite this level of research activity, there is a substantial gap between the number of radiomics-based publications and their actual clinical use [1]. A typical radiomics workflow consists of numerous steps, each of which may be influenced by a variety of factors [2]. This ultimately leads to variability that has a significant impact on reproducibility, and in turn clinical translation, resulting in the well-known "reproducibility crisis" in radiomics [3].

According to the same *PubMed* search, *European Radiology* (435/7,857; 5.5%) is one of the leading journals in terms of the number of radiomics-related publications. Considering all journal categories, it ranks second after *Frontiers in Oncology* (729/7,857; 9.3%) and is followed by *Cancers (Basel)* (309/7,857; 3.9%).

In order to improve the quality of radiomics publications, the *European Radiology* editorial board members in the *Imaging Informatics and Artificial Intelligence* section propose 13 consensus recommendations for future radiomics submissions, which relate to design, data, radiomics methodology, metrics, and reporting of research (Table 1). These are listed briefly below for the authors, but reviewers and readers may also find them helpful.

✉ Burak Kocak
drburakkocak@gmail.com

1 Department of Radiology, University of Health Sciences, Basaksehir Cam and Sakura City Hospital, Basaksehir, Istanbul 34480, Turkey

2 Joint Department of Medical Imaging, University Health Network, University of Toronto, Toronto, ON, Canada

3 The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins Hospital, Baltimore, MD, USA

4 Department of Medicine, Surgery, and Dentistry, University of Salerno, Baronissi, Italy

5 Department of Radiology, St Vincent's University Hospital, Dublin, Ireland

6 Insight Centre for Data Analytics, University College Dublin, Dublin, Ireland

7 Computational Imaging Research Lab, Department of Biomedical Imaging and Image-Guided Therapy, Medical University of Vienna, Vienna, Austria

8 Department of Diagnostic Imaging, National University Hospital, Singapore, Singapore

9 Department of Radiology, University Medical Centre Utrecht, Utrecht University, Utrecht, The Netherlands

10 Centre for Medical Imaging & Centre for Brain Research, Faculty of Medical and Health Sciences, Auckland Bioengineering Institute, University of Auckland, Auckland, New Zealand

11 Department of Radiology, Norfolk & Norwich University Hospital, Norwich, UK

12 Department of Radiology, New York University Grossman School of Medicine, New York, NY, USA

13 Independent Researcher, Handan, Hebei, China

14 Department of Radiology, University Hospital of Cologne, Cologne, Germany

15 Institute for Diagnostic and Interventional Radiology, Goethe-University Frankfurt am Main, Frankfurt, Germany

Table 1 List of ARISE recommendations

Sections	No	Recommendations
Design	1	Have a real-world purpose and unmet need for the radiomics method being proposed
	2	Use and adhere to the checklists, guidelines, and quality scoring tools
Data	3	Split your data into training, validation, and test sets at the beginning of the study with care to prevent information leakage and include a precise description of data partitions
	4	Have internal and/or external test data, use them only once, and never use them for experiments
	5	Be sure your imaging data is heterogeneous with distributional similarity among data partitions
	6	Be sure you have a robust reference standard for the prediction target (i.e., outcome or event) with a similar prevalence to epidemiological data
Radiomics methodology	7	Report parameters of preprocessing steps and radiomics feature extraction process transparently to ensure experiment reproducibility
	8	Ensure values of the radiomics features are stable
	9	Process the radiomics data properly and describe it transparently
	10	Justify the feature set dimensionality considering your sample size
Metrics and reporting	11	Report the predictive performance properly with special consideration for uncertainty estimation
	12	Analyze the clinical usefulness of your radiomics approach supported by the reported study findings
	13	Be transparent with sharing your data, code, and model

Recommendations

Design

1. Have a real-world purpose and unmet need for the radiomics method being proposed.

The objective and unmet need of a radiomics study must be aligned with clinical needs. There should be a robust clinical research question with enough available data to answer the question using radiomics methods and the combined clinical and computer science expertise to apply these methods to answer the question. We encourage submissions that explore the application of radiomics in actual clinical settings, such as the impact on patient management pathways, radiology workflow, and service provision.

2. Use and adhere to the checklists, guidelines, and quality scoring tools.

Following the checklists and guidelines is a simple way to avoid omitting vital details when designing and reporting a study. The use of existing checklists or guidelines in the following references is strongly recommended [4–10]. Among those, CLEAR checklist has recently been developed as a single documentation tool for transparent reporting of radiomics research and endorsed by ESR and EuSoMII [10]. In addition to checklists and guidelines, quality scoring tools can also be utilized [11, 12].

Data

3. Split your data into training, validation, and test sets at the beginning of the study with care to prevent information leakage and include a precise description of data partitions.

The dataset must be split appropriately before performing any action on the imaging data set, such as image and radiomics feature preprocessing and feature dimensionality reduction [13]. Data can be split into training, validation, and test sets. Using a resampling approach for validation such as cross-validation or bootstrapping is highly recommended. The training and validation sets are typically merged in resampling and referred to as development set or simply as training data. Utmost care should be taken to perform a patient-level split. In other words, the data of a particular patient may only appear in exactly one particular data partition. Data partitions must be described with their intended purpose and source. The training and validation sets are used for tuning, experimenting, and model selection. The test set is used to evaluate the generalization performance of the selected model in previously unseen internal and/or external data.

4. Have internal and/or external test data, use them only once, and never use them for experiments.

Whether internal and/or external, unbiased testing of the final model should be performed to properly assess its generalization performance. All iterative experiments must be performed using the training and validation sets alone (i.e., development/construction set). After selecting the best model according to validation metrics (ideally with resampling methods such as cross-validation or bootstrapping), the test data should only be used once to assess its performance. In other words, no alterations to the pipeline should

be performed after obtaining model output on test set data to improve performance. These rules also apply to multiple testing (e.g., performing both internal and external testing or multiple external testing). If any test data is used for experiments, this should be acknowledged with reasons.

5. Be sure your imaging data is heterogeneous with distributional similarity among data partitions.

In order to be more generalizable, data should be heterogeneous enough in terms of demographics (e.g., age, gender, ethnicity), vendors, and imaging protocols, preferably for each data partition and modeling phase. These can be presented with well-designed partition distribution tables. On the other hand, potential sources of model bias should be accounted for to avoid simply learning noise or biased distribution patterns. For example, in the case of an uneven distribution of MRI manufacturers during the training phase, a model may differentiate between manufacturers rather than clinically significant subgroups. Sensitivity and multiparametric analyses should be considered as part of data partitioning to identify the key parameters which should be controlled for during partitioning. Any sort of innate bias within the dataset (e.g., ethnic homogeneity of the local patient population compared to other geographical areas) should be analyzed, acknowledged, and properly discussed if it is present and cannot be avoided.

6. Be sure you have a robust reference standard for the prediction target (i.e., outcome or event) with a similar prevalence to epidemiological data.

Radiomic models should be based on the most reliable reference standards (i.e., outcome or event) that can be reasonably attained. In some cases, using radiology and clinical reports to establish the reference standard or labels may result in models that are not robust. The use of a single expert's evaluation or a system that is susceptible to interobserver variability also represents a risk of introducing bias within a model. Where possible, it is recommended to utilize multiple expert evaluations, consensus evaluations, and interobserver variability-resistant systems (e.g., histopathological grading system). This recommendation is mandatory for the test set, while weaker labeling strategies (such as with clinical reports) may be acceptable when using large-scale datasets to train the model. Furthermore, the prevalence of prediction targets should match epidemiological data, at least for the test set(s).

Radiomics methodology

7. Report parameters of preprocessing steps and radiomics feature extraction process transparently to ensure experiment reproducibility.

All aspects of image processing preceding the extraction of radiomics features should be described in detail to ensure the reproducibility of experiments, as these may have a significant impact on feature values [14]. Gray-level scaling and discretization techniques, de-noising, MRI bias field correction, image interpolation and resampling, motion correction, and thresholding are typical preprocessing techniques. The discretization method must include information about the bin width or bin count, the intended number of gray levels, and the intensity range with minimum bound (particularly useful for the bin width method) [15]. Keeping in mind that the default configuration of the software might not be the optimal choice, care must be taken in setting the discretization parameters (i.e., bin width or bin count) to achieve a reasonable number of gray levels (e.g., 16–128). After preprocessing details, all steps required to reproduce the feature extraction procedure should also be described. In this respect, the following details should be included at a minimum: software-specific information (name, version, and compliance to standards, such as IBSI [16]), 2D or 3D feature extraction, number of features or class names, filters and their parameters, and whether all features are open source.

8. Ensure values of the radiomics features are stable.

When a manual or semi-automatic segmentation method is used, intra- and inter-reader feature reproducibility should be evaluated by performing multiple segmentations of the same region of interest. In the case of fully automated segmentation methods, one should always check the segmentation to ensure accuracy. In contrast to the methods already verified, if a fully automated method is developed as part of the study of interest, a Dice score or Hausdorff distance comparison with ground truth [17] or a further radiomics feature reproducibility analysis (e.g., intraclass correlation) can be performed to ensure algorithm consistency and stability. Furthermore, the stability of the features can also be assessed by perturbing the images, by test-retest imaging analysis if applicable, or by using different software programs.

9. Process the radiomics data properly and describe it transparently.

Typically, radiomics data is processed before modeling. Several algorithms perform better when feature values are in a common scale (i.e., feature scaling), which must be fitted exclusively to the training data [18]. If a significant class or label imbalance exists, techniques such as under-/oversampling or the use of a specific cost function should be considered. Feature selection or data transformations should also be applied to reduce the dimensionality of the data. It is essential to emphasize that steps such as oversampling and

feature selection must be performed using only the training set, with the utmost care to prevent information leakage into the test data, which is also called "double dipping." In the event of cross-validation before the final testing, one must also avoid leakage of relevant information between training and validation folds [19].

10. Justify the feature set dimensionality considering your sample size.

Based on a common rule of thumb, the final number of features for one of the classes should be no more than one-tenth of the number of instances (e.g., patients and tumors) [20]. In case of class imbalance, the least represented class should be considered. For instance, if a study involves 100 instances for each class, the number of features after dimension reduction should not exceed 10. However, this rule of thumb only permits a rough and post hoc evaluation of the dimension and sample size suitability. Although uncommon in radiomics literature, it should be noted that there are additional a priori sample size calculation strategies for multivariable models [21]. The number of features selected should always be appropriately discussed with reference to the sample size.

Metrics and reporting

11. Report the predictive performance properly with special consideration for uncertainty estimation.

Relevant metrics must be provided based on the purpose of the model (i.e., classification or regression), with particular consideration for imbalanced datasets [22]. To evaluate classification models, metrics that capture precision, recall, and true/false positive rates should be reported, such as AUC-ROC (area under the receiver operating characteristic curve), AUC-PR (area under the precision-recall curve), or confusion matrix. If a fixed threshold is set to binarize predictions, the threshold must not be optimized on the test set. To address performance uncertainty, confidence intervals or standard deviations must also be calculated, as appropriate. In the case of multiple models, a statistical comparison of the model performance can be provided, with particular care for multiplicity correction if a frequentist statistical approach is used in comparisons. It is also crucial to note here again that the test set should only be used once for the final performance evaluation after the best model is selected with resampling of training and validation data (e.g., cross-validation).

12. Analyze the clinical usefulness of your radiomics approach supported by the reported study findings.

The potential clinical value of the radiomics approach should be supported by formal analyses within the experimental design. For example, incorporating a non-radiomics strategy for comparison can enable the demonstration of the actual or added value of the proposed method. These may include a widely utilized traditional method in clinical practice (e.g., TNM staging), expert evaluation (radiologists' reading), and clinical or laboratory variables (e.g., PSA values). Comparisons can be based on both predictive performance and clinical utility (e.g., decision curve analysis with an explained clinical rationale).

13. Be transparent with sharing your data, code, and model.

Due to the significance of data, code, and model specifics in the reproducibility of radiomics research, their availability could be used to improve, validate, and disseminate the scientific knowledge generated [11, 23]. Imaging, segmentation, and radiomics feature values are examples of data that can be shared. Code and models should be made available in sufficient detail to enable the replication of experiments. Also, it is relatively simple to wrap a machine learning model into an interface (e.g., by "containerization" of the model) as a ready-to-use system or tool [23]. Providing a containerized model in a public repository allows for an easy independent evaluation of these models without sharing data or re-implementing models. Specifically, the sharing of imaging data is a sensitive issue that must be considered in accordance with institutional and territorial ethical regulations. In contrast, other data types, in addition to code and model, could potentially be made accessible to the public. If data, code, and model cannot be provided, the reason for non-availability must be provided.

Conclusion

Given the disparity between the number of publications and their actual clinical application, we believe this concise consensus guide will be useful to the radiomics community by providing the most essential concepts. In addition, we plan to develop systematic publication appraisal tools with the involvement of the *European Radiology* editorial board in the near future.

Acknowledgements We would like to thank Prof. Yves Menu (Editor-in-Chief of *European Radiology*) and the editorial office for supporting the proposal to prepare this brief guide and for their cooperation.

Funding The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Burak Kocak, MD.

Conflict of interest Daniel Pinto dos Santos is the Deputy Editor of *European Radiology*.
All authors are members of the Imaging Informatics and Artificial Intelligence Section of the Scientific Editorial Board of *European Radiology*. None of the authors has taken part in the review or selection process of this article.
Some authors of this manuscript declare relationships with the following companies: Daniel Pinto dos Santos: Cook Medical (Advisory Board), and Bayer (Speaker Fees).
The other authors of this manuscript declare no relationships with any companies, whose products or services that may be related to the subject matter of the article.

Statistics and biometry Not applicable.

Informed consent Not applicable.

Ethical approval Not applicable.

Study subjects or cohorts overlap Not applicable.

Methodology Opinion based on consensus.

References

1. Pinto Dos Santos D, Dietzel M, Baessler B (2021) A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol 31:1–4. https://doi.org/10.1007/s00330-020-07108-w
2. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö (2019) Radiomics with artificial intelligence: a practical guide for beginners. Diagn Interv Radiol 25:485–495. https://doi.org/10.5152/dir.2019.19321
3. Zwanenburg A (2019) Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging 46:2638–2655. https://doi.org/10.1007/s00259-019-04391-8
4. Vallières M, Zwanenburg A, Badic B, Le Rest CC, Visvikis D, Hatt M (2018) Responsible radiomics research for faster clinical translation. J Nucl Med 59:189–193. https://doi.org/10.2967/jnumed.117.200501
5. Hatt M, Krizsan AK, Rahmim A et al (2023) Joint EANM/SNMMI guideline on radiomics in nuclear medicine : Jointly supported by the EANM Physics Committee and the SNMMI Physics, Instrumentation and Data Sciences Council. Eur J Nucl Med Mol Imaging 50:352–375. https://doi.org/10.1007/s00259-022-06001-6
6. Pfaehler E, Zhovannik I, Wei L et al (2021) A systematic review and quality of reporting checklist for repeatability and reproducibility of radiomic features. Phys Imaging Radiat Oncol 20:69–75. https://doi.org/10.1016/j.phro.2021.10.007
7. Mongan J, Moy L, Kahn CE (2020) Checklist for artificial intelligence in medical imaging (CLAIM): a guide for authors and reviewers. Radiol Artif Intell 2:e200029. https://doi.org/10.1148/ryai.2020200029
8. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK (2020) Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 370:m3164. https://doi.org/10.1136/bmj.m3164
9. Rivera SC, Liu X, Chan A-W et al (2020) Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet Digit Health 2:e549–e560. https://doi.org/10.1016/S2589-7500(20)30219-3
10. Kocak B, Baessler B, Bakas S et al (2023) CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. Insights Imaging 14(1):75. https://doi.org/10.1186/s13244-023-01415-8
11. Lambin P, Leijenaar RTH, Deist TM et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141
12. Cerdá-Alberich L, Solana J, Mallol P et al (2023) MAIC-10 brief quality checklist for publications using artificial intelligence and medical images. Insights Imaging 14:11. https://doi.org/10.1186/s13244-022-01355-9
13. Gidwani M, Chang K, Patel JB et al (2022) Inconsistent partitioning and unproductive feature associations yield idealized radiomic models. Radiology 220715. https://doi.org/10.1148/radiol.220715
14. Wichtmann BD, Harder FN, Weiss K et al (2022) Influence of image processing on radiomic features from magnetic resonance imaging. Invest Radiol. https://doi.org/10.1097/RLI.0000000000000921
15. Duron L, Balvay D, Vande Perre S et al (2019) Gray-level discretization impacts reproducible MRI radiomics texture features. PLoS One 14:e0213459. https://doi.org/10.1371/journal.pone.0213459
16. Zwanenburg A, Vallières M, Abdalah MA et al (2020) The Image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology 295:328–338. https://doi.org/10.1148/radiol.2020191145
17. Reinke A, Tizabi MD, Sudre CH et al (2022) Common limitations of image processing metrics: a picture story. https://doi.org/10.48550/ARXIV.2104.05642
18. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B (2020) Radiomics in medical imaging—"how-to" guide and critical reflection. Insights Imaging 11:91. https://doi.org/10.1186/s13244-020-00887-2
19. Demircioğlu A (2021) Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics. Insights Imaging 12:172. https://doi.org/10.1186/s13244-021-01115-1
20. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR (1996) A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 49:1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3
21. Riley RD, Snell KI, Ensor J et al (2019) Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Stat Med 38:1276–1296. https://doi.org/10.1002/sim.7992
22. Maier-Hein L, Reinke A, Godau P et al (2022) Metrics reloaded: pitfalls and recommendations for image analysis validation. https://doi.org/10.48550/arXiv.2206.01653
23. Kocak B, Yardimci AH, Yuzkan S et al (2022) Transparency in artificial intelligence research: a systematic review of availability items related to open science in radiology and nuclear medicine. Acad Radiol S1076–6332(22):00635–00643. https://doi.org/10.1016/j.acra.2022.11.030