



Reproducibility of radiomics quality score: an intra- and inter-rater reliability study

Tugba Akinci D'Antonoli¹ · Armando Ugo Cavallo² · Federica Vernuccio³ · Arnaldo Stanzione⁴ · Michail E. Klontzas^{5,6} · Roberto Cannella⁷ · Lorenzo Ugga⁴ · Agah Baran⁸ · Salvatore Claudio Fanni⁹ · Ekaterina Petrash¹⁰ · Ilaria Ambrosini⁹ · Luca Alessandro Cappellini¹¹ · Peter van Ooijen¹² · Elmar Kotter¹³ · Daniel Pinto dos Santos^{14,15} · Renato Cuocolo¹⁶ · for the EuSoMII Radiomics Auditing Group

Received: 17 May 2023 / Revised: 3 July 2023 / Accepted: 30 July 2023 / Published online: 21 September 2023
© The Author(s) 2023

Abstract

Objectives To investigate the intra- and inter-rater reliability of the total radiomics quality score (RQS) and the reproducibility of individual RQS items' score in a large multireader study.

Methods Nine raters with different backgrounds were randomly assigned to three groups based on their proficiency with RQS utilization: Groups 1 and 2 represented the inter-rater reliability groups with or without prior training in RQS, respectively; group 3 represented the intra-rater reliability group. Thirty-three original research papers on radiomics were evaluated by raters of groups 1 and 2. Of the 33 papers, 17 were evaluated twice with an interval of 1 month by raters of group 3. Intraclass coefficient (ICC) for continuous variables, and Fleiss' and Cohen's kappa (k) statistics for categorical variables were used.

Results The inter-rater reliability was poor to moderate for total RQS (ICC 0.30–0.55, $p < 0.001$) and very low to good for item's reproducibility ($k = 0.12$ to 0.75) within groups 1 and 2 for both inexperienced and experienced raters. The intra-rater reliability for total RQS was moderate for the less experienced rater (ICC 0.522, $p = 0.009$), whereas experienced raters showed excellent intra-rater reliability (ICC 0.91–0.99, $p < 0.001$) between the first and second read. Intra-rater reliability on RQS items' score reproducibility was higher and most of the items had moderate to good intra-rater reliability ($k = 0.40$ to 1).

Conclusions Reproducibility of the total RQS and the score of individual RQS items is low. There is a need for a robust and reproducible assessment method to assess the quality of radiomics research.

Clinical relevance statement There is a need for reproducible scoring systems to improve quality of radiomics research and consecutively close the translational gap between research and clinical implementation.

Key Points

- Radiomics quality score has been widely used for the evaluation of radiomics studies.
- Although the intra-rater reliability was moderate to excellent, intra- and inter-rater reliability of total score and point-by-point scores were low with radiomics quality score.
- A robust, easy-to-use scoring system is needed for the evaluation of radiomics research.

Keywords Reproducibility of results · Artificial intelligence · Radiomics · Inter-observer variability · Intra-observer variability

Abbreviations

EuSoMII	European Society of Medical Imaging Informatics
GRRAS	Guidelines for Reporting Reliability and Agreement Studies
ICC	Intraclass coefficient
k	Kappa statistics
Q1	First quartal

RQS	Radiomics quality score
TOST	Two one-sided t -tests

Introduction

Radiomics is an analysis tool to extract information from medical images that might not be perceived by the naked eye [1]. Over the course of a decade, several thousand

Extended author information available on the last page of the article

studies have been published spanning diverse imaging disciplines in the field of radiomics research [2]. Nevertheless, the inherent complexity of these advanced methods that are employed to extract quantitative radiomics features may make it difficult to understand all facets of the analysis and evaluate the research quality, let alone to implement these published techniques in the clinical setting [3]. It is evident that easily applicable and robust tools for assessing the quality of radiomics research are needed to move the field forward.

With the aim of improving the quality of radiomics research methods, Lambin et al [4] proposed in 2017 an assessment tool, the radiomics quality score (RQS). Following the ideal workflow of conducting radiomics research, the RQS breaks it down into several steps and aims to standardize them. As a result, the RQS includes 16 items covering the entire lifecycle of radiomics research. Since its introduction in 2017, it has been widely adopted by the radiomics research community, and numerous systematic reviews using this assessment tool have been published [5–9]. However, it can still be inherently challenging for researchers or reviewers to correctly interpret and implement RQS and, therefore, assign scores, which are reproducible; as a result, most of the time the RQS scores are defined with a consensus decision and without a reproducibility analysis in these systematic reviews [5–7, 10–13]. Importantly, no intra- or inter-rater reproducibility analysis was presented in the original RQS publication [4].

According to a recent review article on systematic reviews using the RQS, in most cases the RQS is being used in a consensus approach: 27 out of 44 review articles chose to use consensus scoring, 10 did not even specify how the final scores were obtained, and only 7 of them used intraclass correlation coefficients (ICC) or kappa (k) statistics to assess inter-rater reliability [5]. Despite the positive connotation of a consensus decision, this does not necessarily mean that a score reached by consensus is reproducible. A consensus decision might solely reflect the most experienced rater, as novice voices could be suppressed, resulting in an underestimation of disagreement [14]. The decision to use consensus rather than inter-rater reliability could also presumably be due to challenges in applying the RQS and because ratings cannot be reliably reproduced across raters. Evidently, there is room for improvement in establishing an easily usable and reproducible tool for all researchers.

In this study, we aim to perform a large multireader study to investigate the intra- and inter-rater reliability of the total RQS score and individual RQS items. We believe that a robust method for assessing the quality of radiomics research is essential to carry the field into the future of radiology, rather than ushering in a reproducibility crisis.

Material and methods

The study was conducted in adherence to the Guidelines for Reporting Reliability and Agreement Studies (GRRAS) reporting guidelines [15].

Paper selection

We included studies published recently in *European Radiology*, within an arbitrarily chosen period of 1 month until the start of our study. The following search query is used: (“European Radiology”[Journal]) AND (“radiomics”[Title/Abstract] OR “radiomic”[Title/Abstract]) AND (2022/09/01:2022/10/20[Date—Publication]). *European Radiology* was selected because it is a first-quartile (Q1—Scimago Journal Ranks) journal with the highest number of radiomics publications among all radiology journals; e.g., a PubMed search with keyword “radiomics” or “radiomic” in article title/abstract returns 249 original radiomics articles between January 1, 2021, and December 31, 2022 (Fig. 1).

We only included original research articles and excluded systematic reviews, literature reviews, editorials, letters, and corrections. After applying the inclusion and exclusion criteria, a total of 33 articles were selected for the study, which was above the minimum required sample size, i.e., 30, for the inter-rater reliability studies based on Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research (Fig. 2) [16].

Rater selection and raters’ survey

A total of 9 raters with different backgrounds and experience levels were recruited for the study with an open call within the European Society of Medical Imaging Informatics (EuSoMII) Radiomics Auditing Group. They all completed a survey initially, which was sent to all raters by email to determine their level of expertise in the RQS application as well as the level of expertise in their occupation. Then, they were randomly assigned to the following groups according to their level of expertise: two inter-rater reliability groups, including one with and one without a training session on the use of RQS, and one intra-rater reliability group (Table 1).

The inter-rater reliability group with training (group 1) received a brief training session for the RQS assessment, during which they were instructed by an experienced rater (T.A.D.) about how to rate all items on a random article [17], and then, they separately completed the assessment of all 33 papers. The inter-rater reliability group without

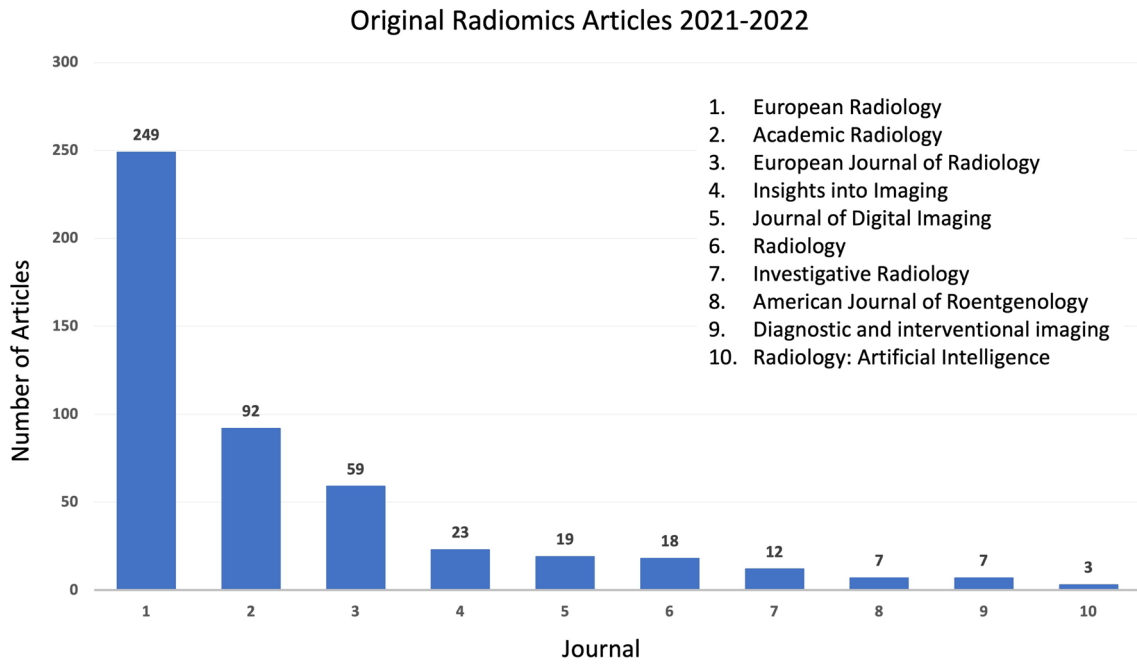


Fig. 1 Bar graphs show the number of original radiomics articles published in first-quarter general radiology journals between 2021 and 2022

Fig. 2 Study flow

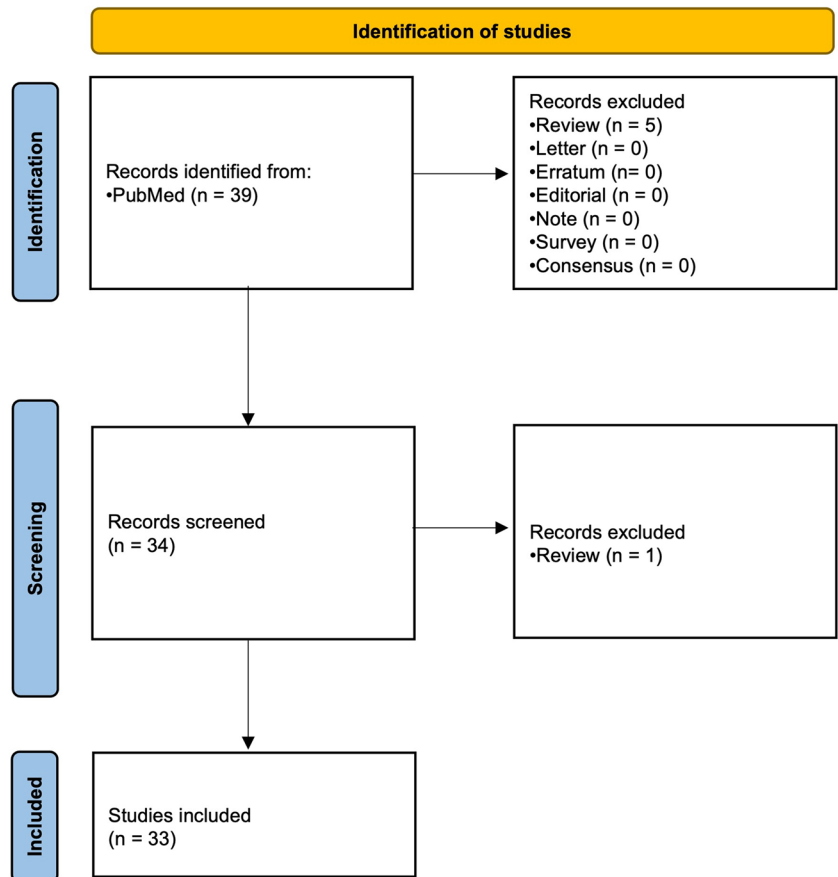


Table 1 Rater characteristics according to the level of RQS rating experience

Rater	RQS rating experience ¹	Group ²	Occupation	Years of experience ³
1 (F.V.)	Novice	2	Radiologist	4
2 (I.A.)	Novice	1	Radiology resident	4
3 (E.A.P)	Intermediate	3	Radiologist	9
4 (S.C.F.)	Intermediate	2	Radiology resident	4
5 (A.B.)	Intermediate	1	Radiologist	8
6 (R.Ca.)	Intermediate	3	Radiologist	3
7 (L.U.)	Advanced	2	Radiologist	5
8 (M.K.)	Advanced	1	Radiology resident	3
9 (A.S.)	Advanced	3	Radiologist	4

¹Novice: I have no previous experience, intermediate: I have some experience with RQS (e.g., 1–2 RQS evaluation), advanced: I have extensive experience with RQS (e.g., 3 or more RQS evaluation)

²Group 1: inter-rater reliability w/ training, group 2: inter-rater reliability w/o training, group 3: intra-rater reliability

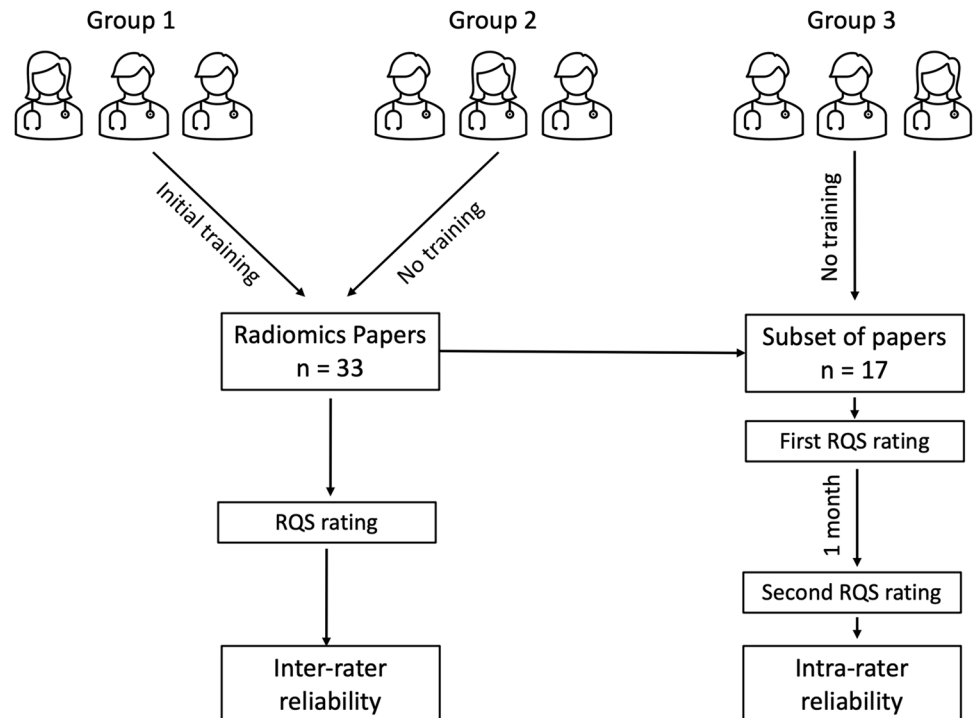
³In occupation

training (group 2) received no training at all on RQS and completed the ratings of all 33 papers. The intra-rater reliability group (group 3) received no training and was asked to score 17 out of 33 selected papers twice 1 month apart to minimize recall (Fig. 3). All raters provided their ratings as they read the article and their available supplementary material. A keyword search was also allowed if needed.

At the end of the study, raters received another survey to investigate the challenges they faced during the RQS assessment and their possible solutions.

Statistical analysis

We used ICC (two-way, single rater, agreement, random effects model) for continuous variables, i.e., total RQS, and Fleiss' and Cohen's *k* statistics for categorical variables, i.e., item scores, as recommended [15, 16, 18]. Cohen's *k* does not support to do comparisons of more than two raters/ratings, and Fleiss' *k* should be used if there are more than two raters/ratings [19]. Therefore, Cohen's kappa is used when there are two ratings/raters, i.e., group 3, and Fleiss' kappa is used when there are more than two ratings/raters, i.e.,

Fig. 3 Study pipeline showing the different groups and their pathways in the study

groups 1 and 2, to compare [19]. We used two one-sided *t*-tests (TOST), a test of equivalence based on the classical *t*-test, to investigate group differences between mean RQS scores [20]. All statistical analysis was carried out with R software (version 4.1.1) and the “irr” and “TOSTER” packages were used [21].

Results

Paper selection

A total of 33 papers were included in this study. Two papers were technical papers, i.e., phantom studies, and all others were original research articles. The characteristics of included studies are shown in Table 2.

Rater selection and raters' survey

Raters were randomly assigned to groups based on the initial survey results (Table 1). After completing the assessments, raters were given another survey to explore the challenges they faced during the RQS assessment and their possible solutions. All responses can be found in Table E1. One of the main problems they faced was the confusion caused by the lack of clear explanations of the RQS items in the main RQS paper and in the RQS checklist [4]. A list of the major issues with RQS along with our recommendations for a simpler approach is presented in Table 3.

Statistical analysis

Inter-rater reliability The inter-rater reliability was poor between raters of group 1 (ICC 0.30; 95% CI [0.09–0.52]; $p=0.0015$), and moderate between raters of group 2 (ICC 0.55; 95% CI [0.29–0.74]; $p<0.001$), and remained low-to-moderate when comparing raters of groups 1 and 2 with the same level of experience (ICC 0.26–0.61). This trend was observed also for intra-group reliability analysis: Raters of group 1 showed poor inter-rater reliability and raters of group 2 moderate inter-rater reliability (Table 4).

Intra-rater reliability In the intra-rater reliability analysis, only rater 3, with intermediate experience level, showed moderate reliability between the first and second read (ICC 0.522; 95% CI [0.09–0.79]; $p=0.009$), whereas rater 6 and rater 9, with advanced experience level, showed excellent intra-rater reliability (ICC 0.91; 95% CI [0.77–0.96]; $p<0.001$ and 0.99; 95% CI [0.96–0.99]; $p<0.001$, respectively).

Reliability of RQS items' score The inter-rater reliability for RQS items' score reproducibility within groups 1 and

2 was very low. The only items that had high inter-rater reliability were items 3 (phantom study) and item 15 (cost-effectiveness analysis). All other items had poor to moderate inter-rater reliability. The intra-rater reliability of RQS items' score was higher and most of the items had moderate to good intra-rater reliability, if not perfect. The mean value and standard deviation of *k* values for group 1 was 0.18 ± 0.33 , for group 2 was 0.43 ± 0.3 , and within group 3 for rater 3 was 0.7 ± 0.3 , rater 6 was 0.75 ± 0.22 , and rater 9 was 0.88 ± 0.27 . Fleiss' *k* for each RQS item of groups 1 and 2 and Cohen's *k* for each RQS item of group 3 are summarized in Table 5.

Moreover, we found that two of the 33 manuscripts included a self-reported RQS which was higher than the scores assigned by the raters in our study as reported in Table 3 [51, 52].

The mean RQS for group 1 was 10.2 ± 3.5 and for group 2 13.2 ± 4 and the mean RQS for group 3 first read was 12.23 ± 5 and second read was 12.4 ± 4.9 (Fig. 4). Two one-sided *t*-tests were applied between the mean RQS value obtained by readers of groups 1 and 2. The lower and upper bounds were calculated to have a statistical power of 0.8 with an alpha of 0.05. Thus, with a lower and upper equivalence bound of ± 2.6 and a mean difference of -3.1 , the *p* value for the lower bound was 0.7 and for the upper bound was <0.001 (Fig. 5).

Discussion

In this study, we conducted a multireader study and investigated the intra- and inter-rater reliability of total RQS as well as individual RQS item scores, involving readers with different experience levels regarding RQS rating. We found that despite being widely adopted, the RQS tool is not straightforward to comprehend and adopt, and its results may not be reproducible in many cases (inter-rater reliability ICC 0.30–0.55, $p<0.001$ and intra-rater reliability ICC 0.522, $p=0.009$ for total RQS; inter-rater group *k*–0.12 to 0.75 and intra-rater group *k*–0.40 to 1 for item's reproducibility). Our results suggest that there is room for improvement to establish an easy-to-use scoring framework for authors, reviewers, and editors to assess the quality of radiomics studies.

To date, RQS has served as a valuable tool to fill the gap for guidance on the quality assessment of radiomics research. Similarly to Lambin et al [4], we believe that the quality of radiomics research should not be compromised, and researchers should transparently report their methods to ensure quality and reproducibility. In addition, to further advance the field, researchers should be incentivized to adopt open science practices. Nonetheless,

Table 2 Characteristics of included papers

Paper	First author	Journal	Publication year	Model utility	Body region	Sample size	Modality	Mean RQS ¹	Self-reported RQS
1	Noortman WA [22]	Eur Radiol	2022	Classification	Abdomen	38	PET-CT	10.3	N/A
2	Bao D [23]	Eur Radiol	2022	Prognostication	Head and neck	216	MRI	16	N/A
3	Chen Q [24]	Eur Radiol	2022	Detection and prognostication	Thorax	240	CT	15.2	N/A
4	von Schacky CE [25]	Eur Radiol	2022	Classification	Musculoskeletal	880	X-ray	13.2	N/A
5	Chu F [26]	Eur Radiol	2022	Prognostication	Abdomen	434	MRI	14.7	N/A
6	Xiang F [27]	Eur Radiol	2022	Prognostication	Abdomen	204	CT	14.8	N/A
7	Zhang H [28]	Eur Radiol	2022	Classification	Abdomen	138	CT	12.7	N/A
8	Zheng Y [29]	Eur Radiol	2022	Classification	Head and neck	388	CT	13	N/A
9	Lin M [30]	Eur Radiol	2022	Detection	Head and neck	489	US	12.5	N/A
10	Jiang J [31]	Eur Radiol	2022	Detection	Neurovascular	403	CT	14.3	N/A
11	Kang JJ [32]	Eur Radiol	2022	Detection	Neuro	149	MRI	8.7	N/A
12	Zhang D [33]	Eur Radiol	2022	Classification	Abdomen	209	MRI	13.7	N/A
13	Ma X [34]	Eur Radiol	2022	Classification	Thorax	612	CT	13.2	N/A
14	Li MD [54]	Eur Radiol	2022	Detection	Technical	108	US	2.5	N/A
15	Xie X [35]	Eur Radiol	2022	Classification	Neuro	89	MRI	8.5	N/A
16	Zhu C [36]	Eur Radiol	2022	Prognostication	Abdomen	106	CT	13.2	N/A
17	Fan Y [51]	Eur Radiol	2022	Detection	Thorax	192	MRI	14.8	27
18	Zhao M [37]	Eur Radiol	2022	Prognostication	Thorax	421	PET-CT	11.2	N/A
19	Frood R [38]	Eur Radiol	2022	Prognostication	Whole body	289	PET-CT	8.5	N/A
20	Zheng Q [39]	Eur Radiol	2022	Detection	Neuro	1650	MRI	8.8	N/A
21	Zhong J [40]	Eur Radiol	2022	Detection and prognostication	Musculoskeletal	144	MRI	14	N/A
22	Cheng B [41]	Eur Radiol	2022	Detection	Thorax	636	CT	14.8	N/A
23	Bi S [42]	Eur Radiol	2022	Prognostication/classification	Head and neck	128	MRI	13	N/A
24	Si N [43]	Eur Radiol	2022	Detection and classification	Cardiovascular-thorax	105	CT	12.3	N/A
25	Eifer M [44]	Eur Radiol	2022	Classification	Thorax	99	PET-CT	6	N/A
26	Chen H [45]	Eur Radiol	2022	Detection and classification	Neuro	609	MRI	11.3	N/A
27	Zhong J [55]	Eur Radiol	2022	Detection	Technical	N/A	CT	4.5	N/A
28	Zhang X [46]	Eur Radiol	2022	Prognostication	Thorax	172	CT	14.2	N/A
29	Zhang H [52]	Eur Radiol	2022	Detection	Neuro	355	CT	12.8	23
30	Zheng YM [47]	Eur Radiol	2022	Prognostication	Head and neck	217	CT	15.8	N/A
31	Salinas-Miranda E [48]	Eur Radiol	2022	Detection	Abdomen	122	CT	9.3	N/A
32	Nagaraj Y [49]	Eur Radiol	2022	Detection	Thorax	2720	CT	14.3	N/A
33	Bleker J [50]	Eur Radiol	2022	Detection	Abdomen	524	MRI	11.7	N/A

¹According to the ratings of 6 raters from groups 1 and 2

any questionnaire or score intended for the evaluation of research or clinical practices should be rigorously evaluated for its reliability and reproducibility. To date, this has not happened for RQS even though it is widely used as a tool to assess the quality of radiomics research. Therefore, we believe that 5 years after its introduction, the

RQS system should be updated to be more easily used by researchers, reviewers, and editors. Recently, a new reporting guideline has been published that covers all requirements, which are necessary to improve radiomics research quality and reliability [53]. We think our recommendations are also in line with this new guideline.

Table 3 The potential reasons for challenges and proposed amendments for radiomics quality score break down by items

Item no	Topic	Score range	Challenges	Recommendation
1	Image protocol quality	+ 1 or + 2	Although this item looks straightforward, it is not exactly clear what authors meant by public protocol; is it an imaging protocol recommended by international guidelines or is it using previously validated protocol?	It is important to explicitly report image protocol to maintain feature reproducibility; however, it should also be defined how much detail will be necessary to ensure quality. We also encourage sharing image protocol explicitly, however, setting a bar for how many details should be presented so that the point could be assigned, and also it should be stated that the imaging protocol should ideally be agreed upon by the radiomics scientific community
2	Multiple segmentation	+ 1	Sole multiple segmentation does not ensure quality; multiple segmentations should always include reproducibility testing to be an added value It is difficult to justify this item since nowadays full automated segmentation algorithms, e.g., nnU-Net, are in use	Manual segmentation is rarely used but if there is multiple segmentation, we recommend assigning a point only if study includes a reproducibility testing We also recommend adapting this item to recent developments and including semi-automatic and automatic segmentation in the rating
3	Phantom study	+ 1	Most of the time, the studies are either phantom study dealing with feature reproducibility or clinical study which does not have any phantom. This item generates a clutter and is not suitable for neither study	We think that quality of phantom studies could be gauged neither by this item nor by total RQS, since most of the time phantom studies will get lower total score as we also observed in our study. Therefore, we recommend removing this item from the scoring system
4	Imaging at multiple time points	+ 1	Most of the times, this was not fulfilled, and even if it is, does not always ensure the quality of the study. Furthermore, this item could easily be misinterpreted and was therefore not reproducible as no clear definition was provided in the checklist as well as in the original RQS article	This item should be clearly explained
5	Feature reduction or adjustment for multiple testing	- 3 or + 3	Giving a minus score is confusing; moreover, it might cause problems while calculating the total score, people could forget using minus, or it could be mistakenly deleted during the analysis	We acknowledge that RQS works with reward and penalize system, but we recommend always assigning a positive score or giving no score at all
6	Multivariable analysis with non-radiomics features	+ 1	This item was confusing and it needs more clarification what non-radiomics features entail	We recommend adding clear explanation about non-radiomics features (i.e., standard of care clinical features or semantic imaging features). It should also be denoted if those non-radiomics features are included in the feature selection process or in the final model. Is the model holistic or if non-radiomics features are completely removed during feature selection?
7	Detect and discuss biological correlates	+ 1	This was the most confusing item according to all raters as most of the researchers superficially claim that their results are correlated with biologic features	We think this item needs clarification
8	Cut-off analyses	+ 1	Even when included in the analysis, the cut-off analysis does not necessarily ensure quality and it brings an arbitrary dichotomization to the table	It is not always necessary to conduct a cut-off analysis

Table 3 (continued)

Item no	Topic	Score range	Challenges	Recommendation
9	Discrimination statistics	+1 or +2	Some raters may be inexperienced	Giving some examples could be helpful
10	Calibration statistics	+1 or +2	If models do not produce probabilistic outputs, then they are automatically penalized	Needs some adjustment for non-probabilistic outputs
11	Prospective study with registration to database	+7	Some studies, although retrospective, include prospectively collected test/validation set. This does not fulfill the requirements of being a prospective study	It should be specified that prospective data collection should be performed explicitly for radiomics clinical trials (i.e., not assigned for retrospective analyses of prospectively collected patients for other studies) and registered in the clinical trial database
12	Validation	-5 to +5	It is very confusing for the raters since there are several steps that should be defined	We think the most important component of a validation is internal and external validation. And internal validation step is sometimes fulfilled with a cross-validation. Therefore, we recommend scoring only these validation steps and keeping the range of the score +1 to +2, one point for each step. Moreover, an independent validation study should be rated the same as external validation
13	Comparison to reference standard	+2	This was one of the most confusing items since most of the studies either does not fulfill this step or there is a lack of clearly defined reference standard	We recommend providing a clear definition of reference standard and outcomes
14	Potential clinical utility	+2	This item is very open ended and most of the time authors tend to mention great potential of clinical utility even though their decision curve analysis shows opposite	We recommend abolishing this item and instead accepting external validation as a sign of potential clinical utility
15	Cost-effectiveness analysis	+1	Almost none of the studies provides this since majority studies are exploratory and there is still lack of prospective studies and RCTs	This item would be unnecessary for usual radiomics studies and might be more suitable for RCTs or radiation oncology studies. We recommend omitting this item
16	Open science and data	+1 to +4	Different incremental steps create a confusion	We think this is very important, but we think scoring system must be simplified. Rating the open data, open code, open model would have been enough
Total points	The total score is often converted into percentage values	-8 to +36	When there are negative results, it is difficult to convert them to percentages. A paper with a score of -8 will score the same in percentage as a paper with a score of 0. So, we can assume that both papers are of low quality, but with a negative score, the magnitude of this "lowness" is difficult to understand Moreover, it is hard to understand when there is no reference point proposed regarding quality	We recommend using only 0 or positive values for scoring. Also, along with the total score we propose implementation of thresholds: 0–25% low, 25–50% average, 50–75% moderate, 75–100% excellent

Table 4 Results of the intra- and inter-rater reliability analysis for overall RQS

	ICC	95% CI	<i>p</i>
Inter-rater analysis			
Overall ICC			
Group 1	0.301	0.09–0.52	0.0015
Group 2	0.549	0.29–0.74	<0.001
Intra-group ICC			
Group 1			
Raters 2 vs 5	0.304	–0.04 to 0.58	0.0418
Raters 2 vs 8	0.232	–0.07 to 0.51	0.0685
Raters 5 vs 8	0.358	0.04–0.61	0.0137
Group 2			
Raters 1 vs 4	0.603	–0.05 to 0.85	0.0398
Raters 1 vs 7	0.510	0.21–0.72	0.001
Raters 4 vs 7	0.529	0.17–0.75	<0.001
Inter-group ICC (matched level of experience)			
Group 1 novice vs group 2 novice	0.255	–0.06 to 0.54	0.0612
Group 1 intermediate vs group 2 intermediate	0.609	0.34–0.79	<0.001
Group 1 advanced vs group 2 advanced	0.349	–0.08 to 0.66	0.0649
Intra-rater analysis			
Group 3			
Rater 3	0.522	0.09–0.79	0.009
Rater 6	0.910	0.77–0.96	<0.001
Rater 9	0.989	0.96–0.99	<0.001

Interestingly, we found slight negativity of the training session that took place prior to the RQS application (according to the two one-sided *t*-test, groups 1 and 2 were not equivalent and statistically different with a lower and

upper equivalence bound of ± 2.6 and a mean difference of -3.1 , lower bound *p* value = 0.7, upper bound *p* < 0.001). The raters of group 1 showed poor inter-rater reliability despite the training and group 2 showed moderate inter-rater

Table 5 Results of the intra- and inter-rater reliability analysis for RQS item reproducibility

	Group 1		Group 2		Group 3					
	<i>k</i> *	<i>p</i>	<i>k</i> *	<i>p</i>	Rater 3		Rater 6		Rater 9	
					<i>k</i> †	<i>p</i>	<i>k</i> †	<i>p</i>	<i>k</i> †	<i>p</i>
Item 1	0.03	0.79	0.32	<0.001	0.46	0.008	0.83	<0.001	–0.03	0.79
Item 2	0.26	0.01	0.51	<0.001	0.57	0.006	0.72	0.002	0.82	<0.001
Item 3	1	0	1	<0.001	1	<0.001	1	<0.001	1	<0.001
Item 4	–0.1	0.24	0.54	<0.001	1	<0.001	1	<0.001	1	<0.001
Item 5	0.0006	0.99	0.46	<0.001	1	<0.001	0.64	0.004	1	<0.001
Item 6	0.35	<0.001	0.55	<0.001	0.41	0.09	1	<0.001	1	<0.001
Item 7	0.38	<0.001	0.19	0.05	0.82	<0.001	0.77	0.001	1	<0.001
Item 8	–0.16	0.11	0.52	<0.001	0.01	0.94	0.36	0.09	1	<0.001
Item 9	0.15	0.06	0.06	0.44	1	<0.001	0.76	<0.001	0.58	0.001
Item 10	0.75	<0.001	0.56	<0.001	0.37	0.04	0.79	<0.001	1	<0.001
Item 11	–0.02	0.83	–0.02	0.83	1	<0.001	1	<0.001	1	<0.001
Item 12	0.22	<0.001	0.43	<0.001	0.59	<0.001	0.80	<0.001	0.89	<0.001
Item 13	–0.04	0.68	0.50	<0.001	0.45	0.02	0.54	0.01	1	<0.001
Item 14	0.23	0.01	0.22	0.02	0.76	<0.001	0.59	0.01	0.85	<0.001
Item 15	–0.01	0.91	1	<0.001	1	<0.001	1	<0.001	1	<0.001
Item 16	–0.21	<0.001	0.03	0.72	1	<0.001	0.30	0.20	1	<0.001

*Fleiss' *k*

†Cohen's

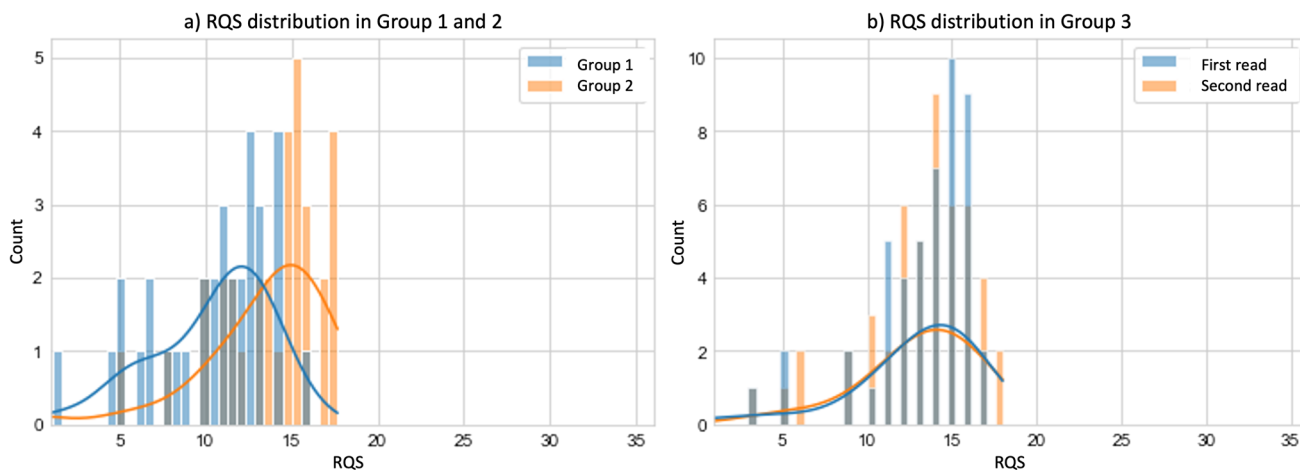


Fig. 4 Histograms and kernel density estimation plots showing the overall distribution of mean RQS separately (a) in group 1 (depicted in blue) and group 2 (depicted in orange) and (b) in group 3 first read (depicted in blue) and second read (depicted in orange)

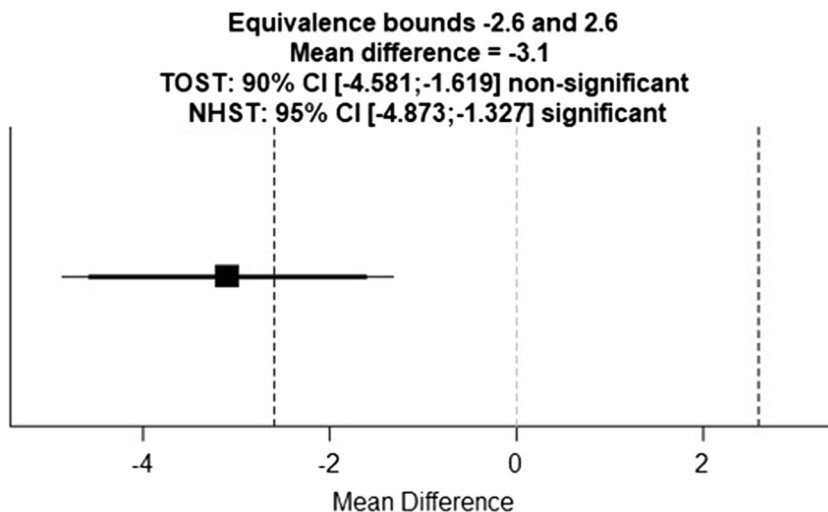
reliability even though they have not received any instructions beforehand. Moreover, we observed the positive effect of more experience only in the intra-rater reliability analysis. The advanced raters showed perfect intra-rater reliability results, whereas the less experienced rater had moderate reliability. We have not observed an effect of experience in the inter-rater reliability analysis.

The raters indicated that the RQS instructions were not self-explanatory in most cases; therefore, they needed more time to interpret the RQS items and consecutively to assign a score. For example, item 4, i.e., “imaging at multiple time points,” was one such item that had low inter-rater reproducibility ($k = -0.1$ in group 1; $k = 0.54$ in group 2) due to unclear item definition in the checklist as well as in the article [4]. It could be argued that this refers to imaging at different time points within the same

examination, i.e., imaging in the arterial/portal venous phase; inspiration/expiration; and test-retest. On the other hand, it could also be argued that this is a hint to longitudinal studies where imaging is performed at different time points, i.e., within 3 months, to perform a delta radiomics analysis. Also, the non-standard range of values, i.e., the sudden change from +1 to +2 to -7 to +7, caused confusion for the authors when assigning a score, without a proper justification of such non-standard range (e.g., for items 5, 12, and 16). A non-standard range would have been acceptable in the case of weighting the item scores according to their importance (Table 3).

One of the problems was that some of the items that may be unusual for the radiology workflow led to confusion instead of clarity. For example, some of the radiomics studies deal only with phantoms with an intention to

Fig. 5 Two one-sided *t*-test graph



cover technical aspects or to test the stability of radiomics features [54, 55]. In this case, an item dealing with phantom studies (item 3) might be a good idea, but in practice, clinical radiomics studies do not necessarily use this phantom step to stabilize their features and do not fulfill this item. Although the transferability of feature robustness from a phantom to a specific biological tissue in the setting of radiomics should still be demonstrated, technically focused phantom studies typically lack clinical validation and therefore tend to achieve lower scores in the RQS system. Similar issues were identified with item 15, which addresses cost-effectiveness analysis. This is very unusual for current radiomics studies, i.e., mostly retrospective, and rarely prospective let alone being included in a randomized controlled study. Also, the definition of cost for radiomics still represents a challenge and, to the best of our knowledge, no published cost-effectiveness analysis for radiomics exists in the literature [56]. Its value in terms of methodological quality could benefit from more research on the topic. Although items 3 and 15 were the most reproducible (Table 5), we argue that they create unnecessary clutter and had a limited impact on overall study quality, as they tended to be always absent (i.e., item 3 and item 15) based exclusively on the study aim or design.

Nowadays, more and more studies utilize deep learning for radiomics analysis; however, the current RQS tool mainly focuses on hand-crafted radiomics, and items specifically addressing the methodological challenges typical to deep learning approaches on radiomics are lacking. Consequently, robust and properly designed deep learning studies might be penalized with a low RQS total score merely because they fail to address questions that are relevant to deep learning methodology. Moreover, in the current RQS tool, sample size analysis or properly selecting the subjects is not rated. We think that sample size analysis and defining the study subjects could be included since study design is one of the most critical steps of a study [57].

We noted that some of the studies included self-reported scores in their publications, but, unfortunately, we found these to be an overly enthusiastic assessment, and observed a large discrepancy when compared with mean RQS results from our multireader analysis [51, 52]. It is not a new phenomenon that researchers tend to overestimate their results and report them within a rose-tinted frame of enthusiasm. This is just a cautionary note for reviewers, editors, and readers to aid correct evaluation of self-reported RQS scores based on our evidence.

Our study had some limitations. We only included a limited amount of papers, but according to the guidelines, it is still more than the minimum required sample size for the inter-rater reliability studies [16]. Moreover, we included articles only from *European Radiology*. However, in the field of

medical imaging, *European Radiology* is the Q1 journal with the highest number of radiomics publications over the past 2 years, ensuring the quality of the studies from a selection of diverse radiomics research areas. In addition, although we intended to explore the effects of training in our study, we did not find any positive effects of training on the reproducibility of RQS. On the one hand, using only one paper as a teaching example might not be sufficient to capture a significant difference. On the other hand, a tool that requires extensive training, even among researchers in the field, to reach adequate reproducibility reveals the limitations of the RQS. Moreover, we have not investigated the effect of training for inter-rater reliability analysis; however, we think the effect of training might be too small to detect as we already found that the intra-rater reliability was moderate to excellent.

In conclusion, we have come a long way in the field of radiomics research, but on the long road to clinical implementation, we need reproducible scoring systems as much as we need reproducible radiomics research. We hope that our recommendations for a more straightforward radiomics quality assessment tool will help researchers, reviewers, and editors to achieve this goal.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00330-023-10217-x>.

Acknowledgements For the EuSoMII Radiomics Auditing Group: Kevin Lipman, Gaia Spadarella, Andrea Ponsiglione, and Anna Andreychenko.

Funding Open access funding provided by University of Basel The authors state that this work has not received any funding.

Declarations

Guarantor The scientific guarantor of this publication is Renato Cuocolo.

Conflict of interest The following authors of this manuscript declare relationships with the following companies: Federica Vernuccio serves as an editorial board member of *European Radiology* and has not taken part in the review and decision process of this paper.

Roberto Cannella received support for attending meetings from Bracco and Bayer; co-funding by the European Union-FESR or FSE, PON Research and Innovation 2014–2020—DM 1062/2021.

Peter van Ooijen: speaker fees from Siemens Healthineers, Bayer, Novartis, and Advisory Board member: MedicalPHIT, ContextFlow. Elmar Kotter: speaker fees from Siemens Healthineers and AbbVie; Advisory Board member ContextFlow, Vienna.

Daniel Pinto dos Santos serves as Deputy Editor of *European Radiology* and has not taken part in the review and decision process of this paper. He received speaker fees from Bayer; Advisory Board member for cook medical.

Renato Cuocolo serves as an editorial board member of *European Radiology* and has not taken part in the review and decision process of this paper.

Federica Vernuccio: Received support to attend meetings from Bracco Imaging S.r.l., and GE Healthcare.

Michail E. Klontzas: Meeting attendance support from Bayer.

The other authors of this manuscript declare no relationships with any companies, whose products or services may be related to the subject matter of the article.

Statistics and biometry Armando Ugo Cavallo (co-author) and Renato Cuocolo (co-author) have significant statistical expertise.

Informed consent Written informed consent was not necessary for this study because no human subjects were involved.

Ethical approval Institutional Review Board approval was not required because no human subjects were involved.

Study subjects or cohorts overlap None.

Methodology

- Retrospective
- Observational
- Multicenter study

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References


1. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: images are more than pictures, they are data. *Radiology* 278:563–577. <https://doi.org/10.1148/radiol.2015151169>
2. Huang EP, O'Connor JPB, McShane LM et al (2022) Criteria for the translation of radiomics into clinically useful tests. *Nat Rev Clin Oncol*. <https://doi.org/10.1038/s41571-022-00707-0>
3. Pinto dos Santos D, Dietzel M, Baessler B (2020) A decade of radiomics research: are images really data or just patterns in the noise? *Eur Radiol* 2–5. <https://doi.org/10.1007/s00330-020-07108-w>
4. 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/nrc1nonc.2017.141>
5. Spadarella G, Stanzione A, Akinici D'Antonoli T et al (2022) Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09187-3>
6. Stanzione A, Gambardella M, Cuocolo R et al (2020) Prostate MRI radiomics: a systematic review and radiomic quality score assessment. *Eur J Radiol* 129:109095. <https://doi.org/10.1016/j.ejrad.2020.109095>
7. Uggla L, Perillo T, Cuocolo R et al (2021) Meningioma MRI radiomics and machine learning: systematic review, quality score assessment, and meta-analysis. *Neuroradiology* 63:1293–1304. <https://doi.org/10.1007/s00234-021-02668-0>
8. Spadarella G, Calareso G, Garanzini E et al (2021) MRI based radiomics in nasopharyngeal cancer: systematic review and perspectives using radiomic quality score (RQS) assessment. *Eur J Radiol* 140:109744. <https://doi.org/10.1016/j.ejrad.2021.109744>
9. Abdurixiti M, Nijiati M, Shen R et al (2021) Current progress and quality of radiomic studies for predicting EGFR mutation in patients with non-small cell lung cancer using PET/CT images: A systematic review. *Br J Radiol*:94. <https://doi.org/10.1259/bjr.20201272>
10. Zhong J, Hu Y, Si L et al (2021) A systematic review of radiomics in osteosarcoma: utilizing radiomics quality score as a tool promoting clinical translation. *Eur Radiol* 31:1526–1535. <https://doi.org/10.1007/s00330-020-07221-w>
11. Wang H, Zhou Y, Li L et al (2020) Current status and quality of radiomics studies in lymphoma: a systematic review. *Eur Radiol* 30:6228–6240. <https://doi.org/10.1007/s00330-020-06927-1>
12. Ursprung S, Beer L, Bruining A et al (2020) Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma—a systematic review and meta-analysis. *Eur Radiol* 30:3558–3566. <https://doi.org/10.1007/s00330-020-06666-3>
13. Kao YS, Te LK (2021) A meta-analysis of computerized tomography-based radiomics for the diagnosis of COVID-19 and viral pneumonia. *Diagnostics* 11. <https://doi.org/10.3390/diagnostic11060991>
14. Bankier AA, Levine D, Halpern EF, Kressel HY (2010) Consensus interpretation in imaging research: is there a better way? *Radiology* 257:14–17. <https://doi.org/10.1148/radiol.10100252>
15. Kottner J, Audigé L, Brorson S et al (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. *Int J Nurs Stud* 64:96–106. <https://doi.org/10.1016/j.jclinepi.2010.03.002>
16. Koo TK, Li MY (2016) A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med* 15:155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
17. Gu D, Hu Y, Ding H et al (2019) CT radiomics may predict the grade of pancreatic neuroendocrine tumors: a multicenter study. *Eur Radiol* 29:6880–6890. <https://doi.org/10.1007/s00330-019-06176-x>
18. Harvey ND (2021) a simple guide to inter-rater, intra-rater and test-retest reliability for animal behaviour studies. *OSF Prepr*:1–13. <https://doi.org/10.31219/osf.io/8stpy>. Accessed at: <https://osf.io/8stpy>
19. McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 22:276–282. <https://doi.org/10.11613/BM.2012.031>
20. Lakens D (2017) Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Soc Psychol Personal Sci* 8:355–362. <https://doi.org/10.1177/1948550617697177>
21. R Core Team (R Foundation for Statistical Computing) (2022) R: A language and environment for statistical computing. <https://www.r-project.org/>
22. Noortman WA, Vriens D, de Geus-Oei LF et al (2022) [18F] FDG-PET/CT radiomics for the identification of genetic clusters in pheochromocytomas and paragangliomas. *Eur Radiol* 32:7227–7236. <https://doi.org/10.1007/s00330-022-09034-5>
23. Bao D, Zhao Y, Li L et al (2022) A MRI-based radiomics model predicting radiation-induced temporal lobe injury in nasopharyngeal carcinoma. *Eur Radiol* 32:6910–6921. <https://doi.org/10.1007/s00330-022-08853-w>
24. Chen Q, Shao JJ, Xue T et al (2022) Intratumoral and peritumoral radiomics nomograms for the preoperative prediction of lymphovascular invasion and overall survival in non-small cell lung cancer. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09109-3>
25. von Schacky CE, Wilhelm NJ, Schäfer VS et al (2022) Development and evaluation of machine learning models based on X-ray

- radiomics for the classification and differentiation of malignant and benign bone tumors. *Eur Radiol* 32:6247–6257. <https://doi.org/10.1007/s00330-022-08764-w>
26. Chu F, Liu Y, Liu Q et al (2022) Development and validation of MRI-based radiomics signatures models for prediction of disease-free survival and overall survival in patients with esophageal squamous cell carcinoma. *Eur Radiol* 32:5930–5942. <https://doi.org/10.1007/s00330-022-08776-6>
 27. Xiang F, Liang X, Yang L et al (2022) Contrast-enhanced CT radiomics for prediction of recurrence-free survival in gallbladder carcinoma after surgical resection. *Eur Radiol* 32:7087–7097. <https://doi.org/10.1007/s00330-022-08858-5>
 28. Zhang H, Meng Y, Li Q et al (2022) Two nomograms for differentiating mass-forming chronic pancreatitis from pancreatic ductal adenocarcinoma in patients with chronic pancreatitis. *Eur Radiol* 32:6336–6347. <https://doi.org/10.1007/s00330-022-08698-3>
 29. Zheng Y, Zhou D, Liu H, Wen M (2022) CT-based radiomics analysis of different machine learning models for differentiating benign and malignant parotid tumors. *Eur Radiol* 32:6953–6964. <https://doi.org/10.1007/s00330-022-08830-3>
 30. Lin M, Tang X, Cao L et al (2022) Using ultrasound radiomics analysis to diagnose cervical lymph node metastasis in patients with nasopharyngeal carcinoma. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09122-6>
 31. Jiang J, Wei J, Zhu Y et al (2022) Clot-based radiomics model for cardioembolic stroke prediction with CT imaging before recanalization: a multicenter study. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09116-4>
 32. Kang JJ, Chen Y, Xu GD et al (2022) Combining quantitative susceptibility mapping to radiomics in diagnosing Parkinson's disease and assessing cognitive impairment. *Eur Radiol* 32:6992–7003. <https://doi.org/10.1007/s00330-022-08790-8>
 33. Zhang D, Cao Y, Sun Y et al (2022) Radiomics nomograms based on R2* mapping and clinical biomarkers for staging of liver fibrosis in patients with chronic hepatitis B: a single-center retrospective study. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09137-z>
 34. Ma X, Xia L, Chen J et al (2022) Development and validation of a deep learning signature for predicting lymph node metastasis in lung adenocarcinoma: comparison with radiomics signature and clinical-semantic model. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09153-z>
 35. Xie X, Yang L, Zhao F et al (2022) A deep learning model combining multimodal radiomics, clinical and imaging features for differentiating ocular adnexal lymphoma from idiopathic orbital inflammation. *Eur Radiol* 32:6922–6932. <https://doi.org/10.1007/s00330-022-08857-6>
 36. Zhu C, Hu J, Wang X et al (2022) A novel clinical radiomics nomogram at baseline to predict mucosal healing in Crohn's disease patients treated with infliximab. *Eur Radiol* 32:6628–6636. <https://doi.org/10.1007/s00330-022-08989-9>
 37. Zhao M, Kluge K, Papp L et al (2022) Multi-lesion radiomics of PET/CT for non-invasive survival stratification and histologic tumor risk profiling in patients with lung adenocarcinoma. *Eur Radiol* 32:7056–7067. <https://doi.org/10.1007/s00330-022-08999-7>
 38. Frood R, Clark M, Burton C et al (2022) Utility of pre-treatment FDG PET/CT-derived machine learning models for outcome prediction in classical Hodgkin lymphoma. *Eur Radiol*:7237–7247. <https://doi.org/10.1007/s00330-022-09039-0>
 39. Zheng Q, Zhang Y, Li H et al (2022) How segmentation methods affect hippocampal radiomic feature accuracy in Alzheimer's disease analysis? *Eur Radiol* 32:6965–6976. <https://doi.org/10.1007/s00330-022-09081-y>
 40. Zhong J, Zhang C, Hu Y et al (2022) Automated prediction of the neoadjuvant chemotherapy response in osteosarcoma with deep learning and an MRI-based radiomics nomogram. *Eur Radiol* 32:6196–6206. <https://doi.org/10.1007/s00330-022-08735-1>
 41. Cheng B, Deng H, Zhao Y et al (2022) Predicting EGFR mutation status in lung adenocarcinoma presenting as ground-glass opacity: utilizing radiomics model in clinical translation. *Eur Radiol* 32:5869–5879. <https://doi.org/10.1007/s00330-022-08673-y>
 42. Bi S, Li J, Wang T et al (2022) Multi-parametric MRI-based radiomics signature for preoperative prediction of Ki-67 proliferation status in sinonasal malignancies: a two-centre study. *Eur Radiol* 32:6933–6942. <https://doi.org/10.1007/s00330-022-08780-w>
 43. Si N, Shi K, Li N et al (2022) Identification of patients with acute myocardial infarction based on coronary CT angiography: The value of pericoronary adipose tissue radiomics. *Eur Radiol* 32:6868–6877. <https://doi.org/10.1007/s00330-022-08812-5>
 44. Eifer M, Pinian H, Klang E et al (2022) FDG PET/CT radiomics as a tool to differentiate between reactive axillary lymphadenopathy following COVID-19 vaccination and metastatic breast cancer axillary lymphadenopathy: a pilot study. *Eur Radiol* 32:5921–5929. <https://doi.org/10.1007/s00330-022-08725-3>
 45. Chen H, Li S, Zhang Y et al (2022) Deep learning-based automatic segmentation of meningioma from multiparametric MRI for preoperative meningioma differentiation using radiomic features: a multicentre study. *Eur Radiol* 32:7248–7259. <https://doi.org/10.1007/s00330-022-08749-9>
 46. Zhang X, Lu B, Yang X et al (2022) Prognostic analysis and risk stratification of lung adenocarcinoma undergoing EGFR-TKI therapy with time-serial CT-based radiomics signature. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09123-5>
 47. Zheng Y-M, Chen J, Zhang M et al (2022) CT radiomics nomogram for prediction of the Ki-67 index in head and neck squamous cell carcinoma. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09168-6>
 48. Salinas-Miranda E, Healy GM, Grünwald B et al (2022) Correlation of transcriptional subtypes with a validated CT radiomics score in resectable pancreatic ductal adenocarcinoma. *Eur Radiol* 32:6712–6722. <https://doi.org/10.1007/s00330-022-09057-y>
 49. Nagaraj Y, de Jonge G, Andreychenko A et al (2022) Facilitating standardized COVID-19 suspicion prediction based on computed tomography radiomics in a multi-demographic setting. *Eur Radiol* 32:6384–6396. <https://doi.org/10.1007/s00330-022-08730-6>
 50. Bleker J, Kwee TC, Rouw D et al (2022) A deep learning masked segmentation alternative to manual segmentation in biparametric MRI prostate cancer radiomics. *Eur Radiol* 32:6526–6535. <https://doi.org/10.1007/s00330-022-08712-8>
 51. Fan Y, Dong Y, Wang H et al (2022) Development and externally validate MRI-based nomogram to assess EGFR and T790M mutations in patients with metastatic lung adenocarcinoma. *Eur Radiol* 32:6739–6751. <https://doi.org/10.1007/s00330-022-08955-5>
 52. Zhang H, Chen H, Zhang C et al (2022) A radiomics feature-based machine learning models to detect brainstem infarction (RMEBI) may enable early diagnosis in non-contrast enhanced CT. *Eur Radiol*. <https://doi.org/10.1007/s00330-022-09130-6>
 53. 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* 14:75. <https://doi.org/10.1186/s13244-023-01415-8>
 54. De LM, Cheng MQ, Da CL et al (2022) Reproducibility of radiomics features from ultrasound images: influence of image acquisition and processing. *Eur Radiol* 32:5843–5851. <https://doi.org/10.1007/s00330-022-08662-1>
 55. Zhong J, Xia Y, Chen Y et al (2022) Deep learning image reconstruction algorithm reduces image noise while alters radiomics features in dual-energy CT in comparison with conventional

- iterative reconstruction algorithms: a phantom study. *Eur Radiol.* <https://doi.org/10.1007/s00330-022-09119-1>
56. Miles K (2020) Radiomics for personalised medicine: the long road ahead. *Br J Cancer* 122:929–930. <https://doi.org/10.1038/s41416-019-0699-8>
57. An C, Park YW, Ahn SS et al (2021) Radiomics machine learning study with a small sample size: single random training-test set split may lead to unreliable results. *PLoS One* 16:e0256152. <https://doi.org/10.1371/journal.pone.0256152>

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

Authors and Affiliations

Tugba Akinci D'Antonoli¹  · **Armando Ugo Cavallo²** · **Federica Vernuccio³** · **Arnaldo Stanzione⁴** · **Michail E. Klontzas^{5,6}** · **Roberto Cannella⁷** · **Lorenzo Ugga⁴** · **Agah Baran⁸** · **Salvatore Claudio Fanni⁹** · **Ekaterina Petrash¹⁰** · **Ilaria Ambrosini⁹** · **Luca Alessandro Cappellini¹¹** · **Peter van Ooijen¹²** · **Elmar Kotter¹³** · **Daniel Pinto dos Santos^{14,15}** · **Renato Cuocolo¹⁶** · **for the EuSoMII Radiomics Auditing Group**

✉ Tugba Akinci D'Antonoli
tugba.akinciantonoli@unibas.ch

¹ Institute of Radiology and Nuclear Medicine, Cantonal Hospital Baselland, Liestal, Switzerland

² Division of Radiology, Istituto Dermopatico dell'Immacolata (IDI) IRCCS, Rome, Italy

³ Institute of Radiology, University Hospital of Padova, Padua, Italy

⁴ Department of Advanced Biomedical Sciences, University of Naples "Federico II", Naples, Italy

⁵ Department of Medical Imaging, University Hospital of Heraklion, Crete, Greece

⁶ Department of Radiology, School of Medicine, University of Crete, Heraklion, Crete, Greece

⁷ Section of Radiology, Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, Palermo, Italy

⁸ MVZ Diagnostikum Berlin GmbH, Diagnostisches Zentrum, Berlin, Germany

⁹ Department of Translational Research, Academic Radiology, University of Pisa, Pisa, Italy

¹⁰ Radiology Department, Research Institute of Children Oncology and Haematology of National Medical Research Center of Oncology n.a.N.N. Blokhin of Ministry of Health of RF, Moscow, Russia

¹¹ Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan, Italy

¹² Department of Radiation Oncology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

¹³ Department of Radiology, University Medical Center Freiburg, Freiburg, Germany

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

¹⁵ Department of Radiology, University Hospital of Frankfurt, Frankfurt, Germany

¹⁶ Department of Medicine, Surgery and Dentistry, University of Salerno, Baronissi, Italy