Diagn Interv Radiol 2024; DOI: 10.4274/dir.2024.242793

ARTIFICIAL INTELLIGENCE AND INFORMATICS

COMMENTARY

Exploring radiomics research quality scoring tools: a comparative analysis of METRICS and RQS

Burak Koçak[1]
Tugba Akinci D'Antonoli[2]
Renato Cuocolo[3]

[1]University of Health Sciences, Başaksehir Çam and Sakura City Hospital, Clinic of Radiology, İstanbul, Türkiye

[2]Cantonal Hospital Baselland, Institute of Radiology and Nuclear Medicine, Liestal, Switzerland

[3]University of Salerno, Department of Medicine, Surgery and Dentistry, Baronissi, Italy

Corresponding author: Burak Koçak

E-mail: drburakkocak@gmail.com

Received 05 April 2024; accepted 25 April 2024.

Epub: 03.05.2024

Publication date:

DOI: 10.4274/dir.2024.242793

Radiomics facilitates the extraction of vast quantities of quantitative data from medical images, which can substantially aid in several diagnostic and prognostic tasks.[1] Although numerous studies have demonstrated promising results with this approach, its integration into clinical practice remains limited, necessitating additional validation for clinical application.[2] A major barrier to this integration is the lack of standardization of key stages in the complex multi-step radiomic pipeline,[3] which could be assessed and enhanced through structured guidelines and quality assessment tools.[4-7]

In 2017, Lambin et al.[8] introduced the radiomics quality score (RQS) as a methodological assessment tool to enhance the quality of radiomics studies. The RQS comprises 16 items that evaluate the entire lifecycle of radiomics research, with a total raw score ranging from −8 to +36. Although the rationale for the scores assigned to each item remains unclear, the radiomics research community has widely adopted this tool since its introduction, leading to numerous systematic reviews.[9] The success of the RQS within the research community also signifies a strong desire for standardization in radiomics, despite its limitations.

Recently, new consensus guidelines specific to radiomics research, namely, the CheckList for EvaluAtion of Radiomics Research (CLEAR) and the METhodological RadiomICs Score (METRICS), have been introduced and endorsed by leading imaging societies.[6,7] CLEAR aims to promote transparent reporting practices, whereas METRICS provides a standardized tool for assessing the methodological quality of radiomics research. METRICS includes 30 items spread over five conditions, designed to accommodate almost all potential methodological scenarios in radiomics research, from traditional handcrafted methods to advanced deep-learning computer vision models.[6] The development process for METRICS involved a modified Delphi method and a broad international panel to mitigate bias and focus on specific aspects of radiomics research related to medical imaging. The European Society of Medical Imaging Informatics has endorsed the METRICS tool, and its website offers an online calculator for the final quality score, which also considers item conditionality (available online at https://metricsscore.github.io/metrics/METRICS.html).[6]

Published in 2024,[6] METRICS is just beginning its journey, and its differences from RQS have not yet been fully explored, which could offer valuable insights for the radiomics community. Therefore, we aimed to compare METRICS and RQS through hypothetical examples, focusing on the unique or missing items of each quality scoring tool. For this comparison, the methodological quality of an ideal hypothetical study was defined as achieving a score of 100% using one tool before being assessed using the other tool, and vice versa. For simplicity, all conditions of METRICS were deemed fulfilled (i.e., scored as "yes") in both comparisons. To establish a baseline, we assumed that a perfect study meets only the minimum requirements of a quality scoring tool (either RQS or METRICS) to attain the highest possible score. This assumption allowed us to evaluate the probable lowest boundary of the highest potential score achievable by the alternative tool. Following the conventions in the literature and recommendations by its developers, the RQS percentage score was calculated by dividing the total points by 36 and multiplying by 100. We also examined the scaling method used for RQS in the literature compared with that of METRICS.

You may cite this article as: Koçak B, Akinci D'Antonoli T, Cuocolo R. Exploring radiomics research quality scoring tools: a comparative analysis of METRICS and RQS. *Diagn Interv Radiol.* 03 May 2024 DOI: 10.4274/dir.2024.242793 [Epub Ahead of Print].

The upper panels of Figure 1 clearly depict a comparison of final quality scores using alternative tools in these hypothetical scenarios. A hypothetical perfect study based on RQS could only achieve a 30% score, which means it lacks up to 70% of the total METRICS percentage score. Conversely, a hypothetical perfect study based on METRICS could reach a 42% score, thus missing 58% of the potential RQS percentage score. Notably, the hypothetical perfect study based on METRICS achieved a higher score in the RQS (42% or 15 total points) compared with the study based on RQS (METRICS: 30%). In the scenario where the perfect study adheres to RQS standards (i.e., RQS: 100%), the requirements for 20 of the 30 items (67%) were not fully met in the METRICS tool. Conversely, in the scenario where METRICS is the standard (i.e., METRICS: 100%), 12 of the 16 (75%) RQS items were not satisfied. Of these, 9 had no direct counterpart in the other tool, whereas the remaining 3 were only partially covered. The lower panels of Figure 1 provide further details about the item-wise comparison in these hypothetical scenarios. Additionally, the items missed in the alternative tools are comprehensively listed in Table 1.

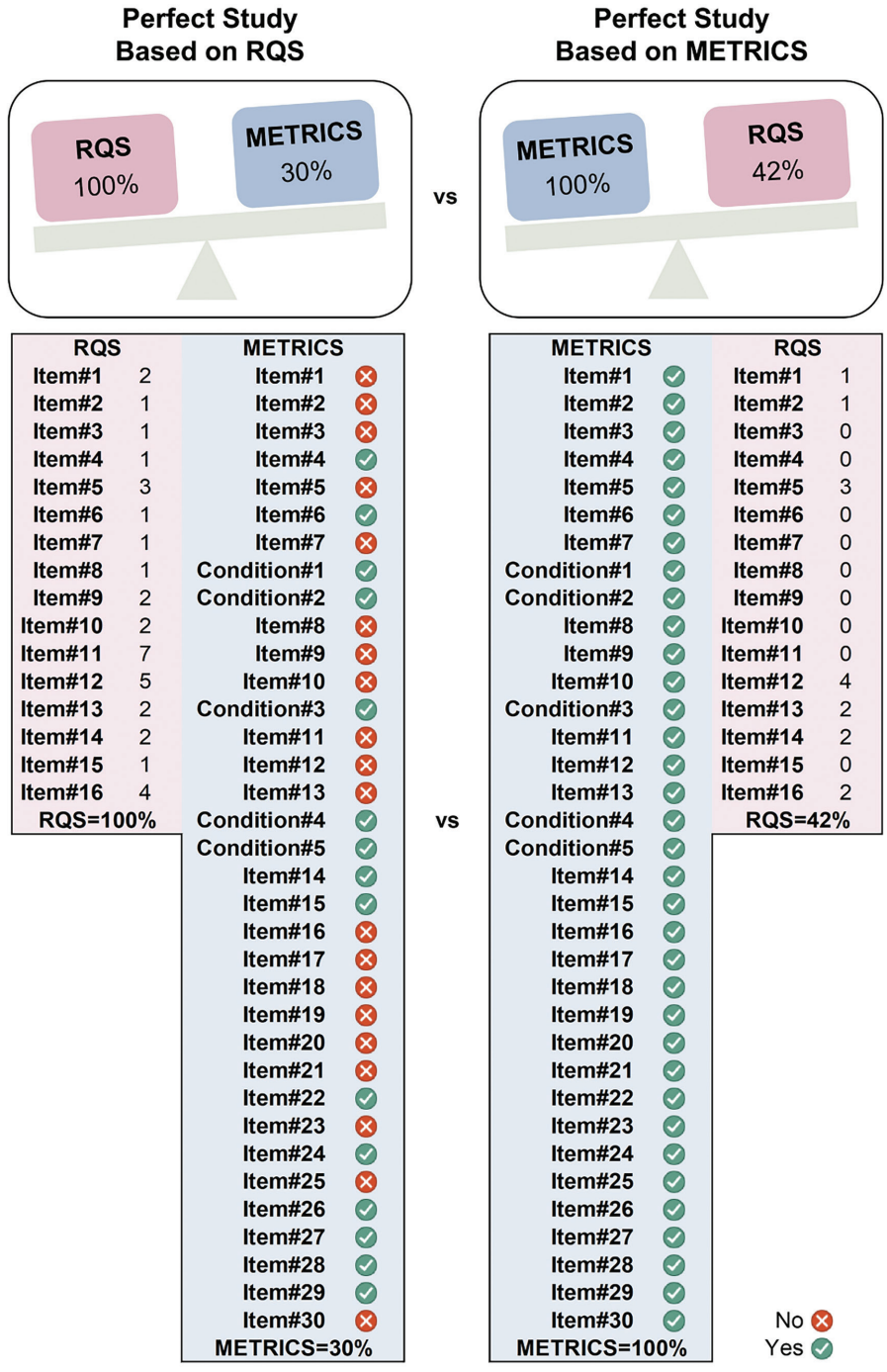

Figure 1. Score-wise (upper panels) and item-wise (lower panels) comparisons of METhodological RadiomICs Score (METRICS) and radiomics quality score (RQS) evaluations in methodologically exemplary hypothetical radiomic studies using RQS and METRICS, respectively. Note: The RQS was calculated by dividing the total score by 36 and then multiplying by 100.

In a perfect study based on RQS, the METRICS evaluation revealed numerous missing items that span almost all sections of the tool, with some sections completely lacking coverage: "study design," "segmentation," "image processing and feature extraction," and "preparation for modeling." The "study design" section of METRICS places substantial emphasis on transparent reporting practices and encourages adherence to specific guidelines tailored to radiomics, such as CLEAR.[7] These METRICS items also highlight crucial aspects of any experimental setup, including the accurate reporting of patient eligibility criteria and reference standards. The "segmentation" section emphasizes the important but often overlooked nuances of data labeling methodology. These include the formal evaluation of fully automatic segmentation (when employed) and the clinical applicability of the segmentation methodology. Specifically, if masks are required for the test set to simulate real-world inference, they should mirror what would reasonably be expected in this context (i.e., produced by a single reader or automated software). "Image processing and feature extraction" considers standardization initiatives such as the Image Biomarker Standardization Initiative, as well as the transparency and appropriateness of settings used in data preprocessing and feature extraction.[5] The items in "preparation for modeling" address key sources of bias, such as proper data partitioning to prevent information leakage during model development and the handling of confounders. Importantly, missed items extend beyond these sections. For instance, METRICS emphasizes the importance of model availability in the "open science" section, which is critical for validating proposed approaches with new data, ideally from a diverse source.

In the same vein, METRICS has not addressed several RQS items. While theoretically possible, certain RQS items such as "phantom study," "multiple time points," "biological correlates," and "prospective study" may be deemed too abstract or lack practical relevance to necessitate their systematic inclusion in every radiomics study.[10] Interestingly, the "prospective study" was initially considered and voted on during the development

Table 1. Missed METRICS and RQS items in the case of methodologically perfect scores in RQS and METRICS, respectively[6,8]

Category	Item no.	Item definition
Missed METRICS items in a perfect study according to RQS[1]	Item#1	Adherence to radiomics and/or machine learning-specific checklists or guidelines
	Item#2	Eligibility criteria that describe a representative study population
	Item#3	High-quality reference standard with a clear definition
	Item#5	Clinical translatability of the imaging data source for radiomics analysis
	Item#7	The interval between imaging used and reference standard
	Item#8	Transparent description of segmentation methodology
	Item#9	Formal evaluation of fully automated segmentation
	Item#10	Test set segmentation masks produced by a single reader or automated tool
	Item#11	Appropriate use of image preprocessing techniques with transparent description
	Item#12	Use of standardized feature extraction software
	Item#13	Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement
	Item#16	Appropriateness of dimensionality compared with data size
	Item#17	Robustness assessment of end-to-end deep learning pipelines
	Item#18	Proper data partitioning process
	Item#19	Handling of confounding factors
	Item#20	Use of appropriate performance evaluation metrics for task
	Item#21	Consideration of uncertainty
	Item#23	Use of uni-parametric imaging or proof of its inferiority
	Item#25	Comparison with simple or classical statistical models
	Item#30	Model availability
Missed RQS items in a perfect study according to METRICS[1]	Item#1	Image protocol quality (2nd sub-item)
	Item#3	Phantom study
	Item#4	Multiple time points
	Item#6	Multi-variable analysis with non-radiomic features
	Item#7	Biological correlates
	Item#8	Cut-off analyses
	Item#9	Discrimination statistics
	Item#10	Calibration statistics
	Item#11	Prospective study
	Item#12	Validation (5th and 6th sub-items)
	Item#15	Cost-effectiveness analysis
	Item#16	Open science and data (any two of 1st, 2nd, or 4th sub-items)

[1]A perfect study is defined as one that meets only the minimum requirements of a quality scoring tool (e.g., RQS or METRICS) to achieve the maximum score available. METRICS, METhodological RadiomICs Score; RQS, radiomics quality score.

of METRICS but failed to reach the consensus threshold for inclusion in the final scoring tool. Likewise, other items were proposed by participants during the METRICS development phase but were excluded from the final tool following open and anonymous discussions throughout the Delphi process, indicating a general consensus on their limited utility. For additional METRICS and RQS items not discussed here, please refer to Table 1.

Although METRICS presents the final score as a percentage value with linear scaling, the RQS does not advocate for this method when converting total RQS points to a percentage. A re-analysis of the papers in the seminal study by Spadarella et al.[9], which included 44 systematic reviews using RQS, revealed that 32 used non-linear scaling (i.e., total points/36*100), and none used linear scaling (i.e., total points + 8/44*100; total points range from −8 to 36, or 44 points). Despite questions about the appropriateness of the non-linear conversion method, this practice follows the developer's suggestion (i.e., 36 = 100%).[8] This method of calculation does not account for negative values in scaling, where both scores of −8 and 0 correspond to 0%, potentially overestimating the score of studies with negative RQS totals. This could lead to the impression that the absence of "feature reduction or adjustment for multiple testing" and "validation" renders the remaining methodological points unsubstantial until an overall positive score is achieved, possibly underestimating the quality of studies on the percentage scale. The upper panel of Figure 2 illustrates a simple comparison of RQS percentage calculations by the widely used non-linear method versus the linear method. The lower panel of Figure 2 shows the impact of using the non-linear method compared with the linear method. This simulation demonstrates that the non-linear method tends to underestimate the final RQS percentage, with a mean, standard de-

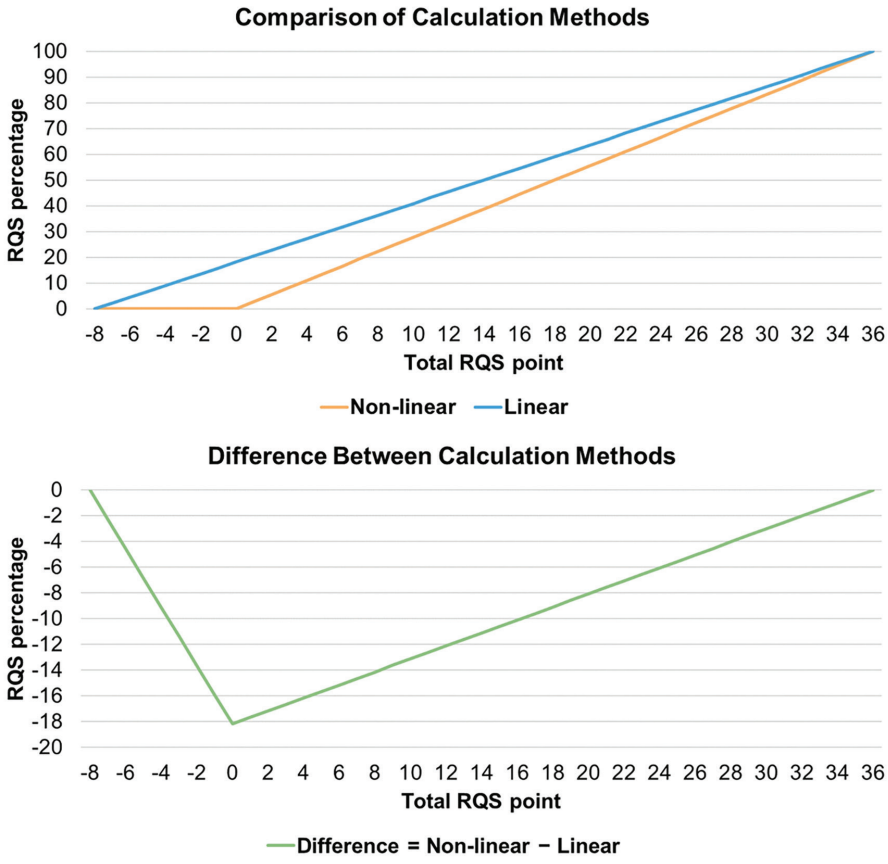

Figure 2. Upper panel: comparison of non-linear (widely used) and linear scaling methods for calculating radiomics quality score (RQS) percentages. Lower panel: differences and consequences resulting from the use of these methods.

viation, and maximum of −8.9%, 5.4%, and 18%, respectively.

In this brief article, we aimed to draw the scientific community's attention to the differences between two quality scoring tools for radiomics research, specifically the recently published METRICS and the well-established RQS. Given the absence of an independent reference standard, which would provide invaluable additional insights, we relied on hypothetical perfect studies to evaluate these tools' relative value and content. Although this approach was hypothetical, it underscored the distinct focus of each tool on different aspects of the radiomic pipeline, given the substantial disparity in relative scores and missed items. Therefore, a direct comparison of the scores from these tools is not feasible, and researchers should consider the unique features of each tool. Based on the insights from this analysis and the emerging limitations regarding the reproducibility and accuracy of the RQS percentage score,[9,10] METRICS may be the preferable choice if only one tool is to be used.

Conflict of interest disclosure

Burak Koçak, MD, is Section Editor in Diagnostic and Interventional Radiology. He had no involvement in the peer-review of this article and had no access to information regarding its peer-review. Burak Koçak, Tugba Akinci D'Antonoli, and Renato Cuocolo took part in the development of METRICS.

References

1. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging*. 2020;11(1):91. [CrossRef]
2. Zhong J, Lu J, Zhang G, et al. An overview of meta-analyses on radiomics: more evidence is needed to support clinical translation. *Insights Imaging*. 2023;14:111. [CrossRef]
3. Cobo M, Menéndez Fernández-Miranda P, Bastarrika G, Lloret Iglesias L. Enhancing radiomics and Deep Learning systems through the standardization of medical imaging workflows. *Sci Data*. 2023;10(1):732. [CrossRef]
4. Whybra P, Zwanenburg A, Andrearczyk V, et al. The image biomarker standardization initiative: standardized convolutional filters for reproducible radiomics and enhanced clinical insights. *Radiology*. 2024;310(2):e231319. [CrossRef]
5. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. 2020;295(2):328-338. [CrossRef]
6. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al. METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII I. *Insights Imaging*. 2024;15(1):8. [CrossRef]
7. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75. [CrossRef]
8. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-762. [CrossRef]
9. Spadarella G, Stanzione A, Akinci D'Antonoli T, et al. Systematic review of the radiomics quality score applications: an EuSoMII Radiomics Auditing Group Initiative. *Eur Radiol*. 2023;33(3):1884-1894. [CrossRef]
10. Akinci D'Antonoli T, Cavallo AU, Vernuccio F, et al. Reproducibility of radiomics quality score: an intra- and inter-rater reliability study. *Eur Radiol*. 2024;34(4):2791-2804. [CrossRef]